

Supplementary Materials for Learning When to Look: A Disentangled Curriculum for Strategic Perception in Multimodal Reasoning

Siqi Yang^{1*}, Zilve Gao^{2*}, Haibo Qiu¹, Fanfan Liu¹, Peng Shi¹, Zhixiong Zeng¹,
Qingmin Liao², Lin Ma^{1†}

¹Meituan, ²Tsinghua University

siqi.yang@uq.net.au, gaozl24@mails.tsinghua.edu.cn, haibo-qiu@outlook.com
liufanfan@mail.ucas.ac.cn, pengshi.scholar@gmail.com, zengzhixiong@meituan.com
liaoqm@tsinghua.edu.cn, forest.linma@gmail.com

In the following supplementary material, we first show more implementation details of our proposed method, including training (Section A), pivotal reasoning keyword (Section B) and perception-grounded data generation (Section C). Finally, we demonstrate additional qualitative examples in Section D.

A. More Implementation Details

In the SFT stage, we utilized the LlamaFactory training framework [8]. For text-only SFT, we used the GRAMMAR dataset [1], which contains 47K samples focused on reasoning tasks. For the multimodal SFT phase, we collected and integrated open-source datasets, such as ARES [2] and CLEVR-Math [4]. Our ablation study (Section 4.5.1), based on Qwen2.5VL as the base model, demonstrated that performing text-only cold start followed by multimodal SFT enables the model to develop stronger cognitive behaviors in vision-language tasks. Based on this finding, we further trained the text-only cold-started model on our proprietary PG-COT dataset until the loss converged below 0.1, resulting in a cold-start model capable of autonomously inserting visual perception segments. In the reinforcement learning stage, we employed the EasyR1 framework [6] and the Vir139K training dataset [5]. The model was optimized using the DAPO algorithm [7] combined with our proposed composite reward function, and training was continued until the accuracy metric converged. The specific training parameters for both SFT stages (which share the same configuration) and the RL stage are summarized in Table 6.

Table 6. Key Hyperparameters for Training Stages

Component	Hyperparameter	Component	Hyperparameter
Cold Start	Learning Rate = 5.0×10^{-5}	DAPO	Max Response Length = 8192
	Gradient Accumulation = 8		Rollout Batch Size = 512
	Number of Epochs = 5		Actor Global Batch Size = 128
	LR Scheduler = Cosine		Learning Rate = 1.0×10^{-6}
	Warmup Ratio = 0.05		Weight Decay = 1.0×10^{-2}
	Max Sequence Length = 32768		Optimizer Strategy = adamw
	Precision = BF16		Generation Temperature = 1.0
DeepSpeed = Zero2	Generation Top-p = 1.0		

B. Details for Pivotal Reasoning Keywords

We propose a pivotal perception reward to train the model’s policy on strategic perception timing, based on the hypothesis that visual grounding should be triggered by linguistic markers of cognitive uncertainty. The reasoning-pivotal keywords in \mathcal{K} are shown in Figure 4.

*Equal contribution. Work done when Zilve Gao worked as an intern with Meituan.

†Corresponding author.

Pivotal Reasoning Keywords

Metacognitive Markers:

let me recalculate, let's revise, let me reconsider, let's rethink, let me check again, double-check, verify, re-evaluate, reassess, revisit, let's try again, wait, hold on, however, but, although, on second thought, wait a second, is that right, did I miss something, I'm not sure, I am not sure, I wonder, perhaps, maybe, I think, mistake, error, incorrect, wrong, I was mistaken, that's not right, this contradicts, I need to reconsider, doesn't make sense, I made an error, contradiction, flaw, invalid

Structural Transitions:

first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, step, steps, step 1, step 2, option, options, approach, approaches, method, methods, way, ways, process, procedure, stage, stages, phase, phases, part, parts, section, sections, aspect, aspects, next, then, after, following, subsequently, initially, finally, lastly, ultimately, sequence, order, progression

Visual Re-Examination Cues:

observe, observing, observation, observations, notice, noticing, noticed, see, seeing, saw, seen, look, looking, looked, examine, examining, examined, examination, check, checking, checked, identify, identifying, identified, identification, recognize, recognizing, recognized, recognition, detect, detecting, detected, detection, find, finding, found, discover, discovering, discovered, discovery, analyze, analyzing, analyzed, analysis, consider, considering, considered, consideration, assess, assessing, assessed, assessment, inspect, inspecting, inspected, inspection, review, reviewing, reviewed, investigate, investigating, investigated, investigation, explore, exploring, explored, exploration

Figure 4. Pivotal reasoning keywords.

C. Details for Perception Grounded CoT Data Generation

We transform existing multimodal reasoning datasets into PG-CoT format using a teacher MLLM, *i.e.*, Seed1.5-VL [3]. The data generation pipeline operates in three steps: (1) analyze the original CoT to identify logical breakpoints where visual evidence is required, (2) generate fine-grained observations anchored to specific visual elements, and (3) insert perception segments while preserving reasoning fidelity. The prompts for these three steps are as follows:

Prompt for Image Description Generation

Please provide an extremely detailed and comprehensive description of this image in English.

Requirements:

1. Describe ALL visible text content (including titles, body text, numbers, formulas, labels, etc.)
2. Describe ALL visual elements (shapes, charts, symbols, photos, illustrations, etc.), including their colors, positions, and size relationships
3. Describe the spatial layout and arrangement of elements
4. If there are tables, charts, or other structured information, describe their content in detail
5. Be as thorough as possible so that a reader could understand the complete content of the image based solely on your description

Please start describing directly without using numbered points or specific formatting. Use natural, flowing language to present all the information in the image comprehensively.

Prompt for Perception Location Recommendation

You are an expert in multimodal reasoning and data annotation.

Your task is to review a text-only Chain-of-Thought (CoT) reasoning process for a question that originally included an image. You need to:

1. Suggest where perception segments (i.e., visual observations) could be inserted in the reasoning, and specify what type of visual information should be described in each segment.
2. Identify any sentences or phrases within the reasoning that already describe visual information, and recommend wrapping these with `<perception>...</perception>` tags.

Original Question:

{Insert the original question here}

Thinking Content (MUST be preserved exactly as shown below):

{Insert the original CoT reasoning here}

Important Context:

The original question was accompanied by an image. The reasoning provided here is based only on text, but in a multimodal setting, certain reasoning steps could be enhanced by directly grounding them in visual evidence.

Core Instructions:

1. Carefully read the entire CoT reasoning to understand its logic and structure.
2. For each reasoning step, ask: "Would this step benefit from being grounded in visual evidence from the image?"
3. Suggest specific points (only after punctuation marks) where a perception segment could be inserted to support or guide the reasoning.
4. For each suggested perception segment, clearly specify what type of visual information should be described (e.g., chart data, labels, object attributes, spatial relationships, mathematical expressions, symbols, text, numbers, geometric shapes, etc.).
5. Carefully examine the reasoning content and identify any sentences or phrases that already directly describe visual information from the image. For these, recommend wrapping them with `<perception>...</perception>` tags.

Continued on next page ...

Prompt for Perception Location Recommendation (Cont.)

6. Do NOT generate actual perception content—only indicate where and what should be described, and which existing segments should be wrapped, as if preparing instructions for a multimodal annotator.
7. Your suggestions must be based entirely on the reasoning process and the original question. Do not invent or assume visual details that are not logically implied or necessary.
8. Do NOT remove, rewrite, or summarize any part of the original reasoning. Your job is to annotate the reasoning process, not to improve or alter it.

Output Format:

- For each suggested insertion point, list:
 1. The exact location (e.g., after which sentence or phrase).
 2. The type of visual information that should be described in the perception segment (in clear, concise language).
- For each existing sentence or phrase that describes visual information, specify its location and recommend wrapping it with `<perception>...</perception>` tags.
- Maintain the order and structure of the original reasoning.
- When specifying the type of visual information, use the same language as the original content (e.g., use Chinese if the original is in Chinese, English if in English).

Rules for Perception Segment Suggestions:

- Only suggest perception insertions after logical break points (punctuation marks, not in the middle of sentences or formulas).
- The suggested visual information must be directly relevant to the reasoning step and necessary for answering the main question.
- Do not suggest irrelevant or distracting visual observations.
- Distribute suggestions evenly throughout the reasoning trace, where they would add value.
- If a reasoning step does not require visual grounding, do not suggest a perception segment.
- Do not suggest insertions in the middle of mathematical formulas or equations.
- For existing visual description sentences, only recommend wrapping those that are directly and factually describing the image.

Examples of CORRECT suggestions:

- After “We need to solve for the range.”
 - Suggested perception: Describe the chart showing export data for four years.
- After “Looking at the chart data: 2004 shows \$8.66 billion, 2005 shows \$10.92 billion, 2006 shows \$13.41 billion, 2007 shows \$16.7 billion.”
 - Suggested perception: Describe the exact values and labels visible in the chart.
- Sentence: “The image shows a right triangle with sides labeled 3, 4, and 5.”
 - Recommendation: Wrap this sentence with `<perception>...</perception>` tags.

Examples of INCORRECT suggestions:

- In the middle of a sentence or formula.
- After reasoning steps that do not rely on visual information.
- Suggesting perception segments that are not relevant to the question.
- Wrapping sentences that do not actually describe visual evidence.

Begin your output now:

Prompt for Perception Segment Insertion

You are an expert in multimodal reasoning and data annotation.

Your task is to enhance a text-only Chain-of-Thought (CoT) reasoning for a question that originally included an image.

You will use the provided suggestions to:

1. Wrap existing sentences or phrases that directly describe visual information with `<perception>...</perception>` tags, according to the suggestions.
2. Insert new perception segments at the recommended locations, following the suggestions on what type of visual information to describe.
3. Ensure all perception segments are concise, factually accurate, and based only on what can be directly observed in the image.
4. For each perception segment, cross-check with the provided image description for consistency, but rely primarily on your own observation of the image.
5. If the suggested visual evidence is not present in the image, simply skip inserting that perception segment.

Original Question:

{Insert the original question here}

Original Image:

{Insert the image here or provide the image file}

Image Description:

{Insert the pre-generated image description here}

Thinking Content (MUST be preserved exactly as shown below):

{Insert the original CoT reasoning here}

Perception Suggestions:

{Insert the list of suggestions, including:

- Locations for inserting new perception segments and what to describe
- Locations of existing visual description sentences/phrases to be wrapped}

Core Instructions:

1. Carefully observe the provided image to understand all visual elements present, including: natural images, photographs, charts, graphs, diagrams, mathematical expressions, formulas, text content, numbers, symbols, geometric shapes, and any other visible elements.
2. Read the provided Chain-of-Thought reasoning and the perception suggestions.
3. For each suggestion to wrap an existing sentence or phrase, locate it in the reasoning and wrap it with `<perception>...</perception>` tags.
4. For each suggestion to insert a new perception segment, add a `<perception>...</perception>` block at the specified location (only after punctuation), describing the recommended type of visual information observed in the image.
5. For each perception segment, cross-check with the image description for consistency, but base your final judgment on your own observation.
6. If the suggested visual evidence is not present in the image, simply skip inserting that perception segment.
7. Each perception segment must contain a single, concise, factually accurate sentence that is directly verifiable from the image.

Continued on next page ...

Prompt for Perception Segment Insertion (Cont.)

8. Do not invent or assume details not visible in the image. Base all perception content on direct observation.
9. Do not remove, rewrite, or summarize any part of the original reasoning. Preserve the logical structure and meaning.
10. Use the same language as the original reasoning when writing perception segments.
11. Insert perception segments only at logical break points (after punctuation marks, never in the middle of sentences or formulas).

Output Format:

- Return the enhanced reasoning content, with `<perception>...</perception>` tags correctly wrapped around existing visual description sentences/phrases, and new `<perception>` segments inserted at the recommended locations.
- Each `<perception>` segment should be concise and directly describe what is observable in the image.

Examples of CORRECT placement:

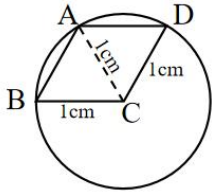
"`<perception>`The image shows a right triangle with legs of length 3 and 4`</perception>` According to the Pythagorean theorem, we need to calculate the hypotenuse length."

"`<perception>`The bar chart shows sales data for three different products, with Product A as blue bars, Product B as red bars, and Product C as green bars`</perception>`We need to compare the sales performance of these three products."

Begin your output now:

D. More Case Studies

We provide three illustrative prediction examples, contrasting our approach with the Qwen2.5-VL-DAPO baseline. Examination of our reasoning trajectories reveals that perception segments are systematically invoked following pivotal reasoning keywords, validating the efficacy of our reward design in learning strategic perception timing.



Question: As shown in the figure, a circle is drawn with center C, and points A and B are exactly on the circle. The lengths of BC, AC, and CD are as shown in the figure, and $AB = CD$. What is the measure of $\angle ABC$?
A. 45° ; B. 60° ; C. 72° ; D. 90° ; E. No correct answer

Answer (Qwen2.5-VL-DAPO):

<think>

To determine the measure of $\angle ABC$, we need to recognize that since $AB = CD$ and all line segments are given as 1 cm, quadrilateral ABCD is a rhombus (as all sides are equal). Given that C is the center of the circle, triangle ABC and triangle ADC are isosceles triangles with $AC = BC = CD = AD = 1$ cm.

Since $AB = CD = 1$ cm, quadrilateral ABCD is a rhombus. Because it lies inscribed in a circle (a cyclic quadrilateral), it must be a rectangle (as opposite angles sum to 180° and all sides are equal implying right angles at vertices A and C).

Thus, $\angle ABC$ must be a right angle (90°) because in a cyclic quadrilateral where opposite sides are equal and inscribed in a circle, it forms a rectangle implying right angles at vertices.

Therefore, $\angle ABC = 90^\circ$.

</think>

\boxed{D} (WRONG)

Answer (Ours):

<think>

Okay, let's try to solve this problem. **First, let's look at the diagram.** <perception>There's a circle with center C, so points A and B are on the circle, which means CA and CB are radii</perception>. The problem says BC, AC, and CD are as shown. The lengths: BC is 1 cm, AC is 1 cm, CD is 1 cm. Oh, so $CA = CB = CD = 1$ cm. So triangle ABC has $CA = CB = 1$ cm, and $AB = CD$. But CD is also 1 cm, so $AB = 1$ cm. Wait, so triangle ABC has sides $CA = CB = AB = 1$ cm. That makes triangle ABC equilateral. So all angles are 60 degrees. Therefore, angle ABC is 60° , which is option B.

Let me double-check. <perception>The diagram clearly labels CA, CB, and CD as 1cm each, and AB is equal in length to CD</perception> The center is C, so CA and CB are radii, so $CA = CB = 1$ cm. CD is also 1 cm, so $AB = CD = 1$ cm. So triangle ABC has three sides of 1 cm, so it's equilateral. Therefore, angle ABC is 60° , which is option B. **That seems right.** <perception>The geometric arrangement confirms triangle ABC has all sides equal, identifying it as an equilateral triangle</perception> Maybe I'm missing something? Let me think again. The problem states $AB = CD$, and CD is 1 cm, so AB is 1 cm. CA and CB are both 1 cm, so triangle ABC is equilateral. So angle ABC is 60° , which is option B. So the answer should be B. Wait, but let me check the options again. Option B is 60° , so boxed B. So **\boxed{B}**.

</think>

To solve the problem, we analyze the given diagram and information:

<perception>The circle has center C, so CA, CB, and CD are radii. From the diagram, $CA = CB = CD = 1$ cm.</perception>

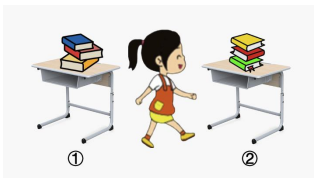
AB is given to be equal to CD, so $AB = 1$ cm.

In triangle ABC, $CA = CB = AB = 1$ cm, making triangle ABC equilateral.

In an equilateral triangle, all angles are 60° , so $\angle ABC = 60^\circ$.

Thus, the measure of $\angle ABC$ is 60° , corresponding to option B.

\boxed{B} (CORRECT)



Question: As shown in the diagram, the color of the book at the bottom of the desk in front of Lucy is ()

A. Red; B. Green; C. Blue; D. Yellow; E. No correct answer

Answer (Qwen2.5-VL-DAPO):

<think> To determine the color of the book at the bottom of the desk in front of Lucy, I need to identify which desk is in front of her. Since the girl is positioned between the two desks, the desk on the left (labeled 1) is in front of her. The books on desk 1 (from left to right) are:

Yellow

Red

Blue

Therefore, the book at the bottom of desk 1 is yellow.

Thus, the correct answer is D. **</think>**

\boxed{D} (WRONG)

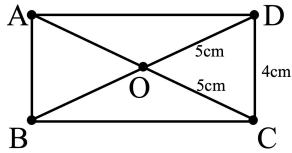
Answer (Ours):

<think> Okay, let's try to figure this out. The question is about the color of the book at the bottom of the desk in front of Lucy. **First, I need to look at the diagram.** **<perception> The diagram has two desks: desk on the left and desk on the right. Lucy is between them. The desk in front of her is desk 1, because she's facing desk 1. </perception>** Now, let's check the books on desk 1. The books there are stacked: yellow, red, green? Wait, no, wait. Wait, the books on desk 1 are a stack. **Let me see: <perception>the top book is yellow, then red, then green</perception>**. So the bottom book is green. Let me consider the options, the options are A. Red, B. Green, C. Blue, D. Yellow. So the bottom book is green, which is option B. **But I need to confirm the picture, <perception>The image shows desk 1 with a stack of books where the bottommost book is green</perception>** So the answer should be B. Wait, but wait, maybe I'm misunderstanding. Wait, the desk in front of Lucy—since Lucy is walking towards desk 1, the desk in front of her is desk 1. So the books on desk 1: the bottom one is green. So option B. That's correct. **Wait, but let me make sure. <perception>Both desks have stacks of books, and desk 1's bottom book is green as observed in the image</perception>** So the answer is B. So boxed B. The options are A to E. So yes, B. **</think>**

To determine the color of the book at the bottom of the desk in front of Lucy, we first identify which desk is in front of her. **<perception>Lucy is positioned between two desks, and the desk facing her is desk 1. On desk 1, the books are stacked with the bottom book being green. </perception>**

The options are A. Red, B. Green, C. Blue, D. Yellow. The correct answer corresponds to the green book at the bottom.

\boxed{B} (CORRECT)



Question: As shown in the figure, the diagonals of rectangle ABCD are AC and BD. What is the length of OC?
 () cmA. 3; B. 4; C. 5; D. 6; E. No correct answer

Answer (Qwen2.5-VL-DAPO):

<think>

To determine the length of OC in rectangle $ABCD$, we start by noting that the diagonals of a rectangle are equal and bisect each other. Therefore, $AC = BD$ and $AO = OC = BO = OD$. Given: $BD = 5$ cm. Since BD is bisected at point O , it follows that: $BO = OD = \frac{BD}{2} = \frac{5}{2} = 2.5$ cm. Thus, $OC = 2.5$ cm. However, we need to find OC in terms of the full length of diagonal components: Since $AC = BD = 5$ cm and $OC = AO = \frac{AC}{2}$. Now, using the Pythagorean theorem in $\triangle DOC$: $AC^2 = AD^2 + DC^2$. Substituting the given lengths: $AC = \sqrt{AD^2 + DC^2} = \sqrt{4^2 + 5^2} = \sqrt{16 + 25} = \sqrt{41}$. But since OC is half of diagonal: $OC = \frac{AC}{2} = \frac{5}{2} = 2.5$ Thus, $OC = 2.5$ Therefore, the length of OC is 2.5.

</think>

\boxed{D} (WRONG)

Answer (Ours):

<think>

Okay, let's see. The problem is about a rectangle ABCD with diagonals AC and BD intersecting at point O. We need to find the length of OC. First, I remember that in a rectangle, the diagonals are equal and they bisect each other. So, $AC = BD$, and O is the midpoint of both diagonals. That means $AO = OC = BO = OD$. Wait, <perception>the diagram shows that one of the segments is 5 cm</perception>. Wait, which one? Oh, the diagonal BD is labeled 5 cm? Wait, no, the segment OD is labeled 5 cm? Wait, looking at the diagram: <perception>the diagonal BD is split into BO and OD</perception>, so if O is the midpoint, then $BD = 2 * OC$. Wait, but the problem says "the diagonals of rectangle ABCD are AC and BD". So AC and BD are the diagonals, and they intersect at O, so O is the midpoint. So OC is half of AC, but also, since diagonals are equal, $AC = BD$. Let me check again. <perception>The image shows the segment OD labeled "5cm"</perception> Since in a rectangle, diagonals bisect each other, so $OC = OD$. Therefore, OC is 5 cm. So the answer is C. But wait, let me make sure. <perception>The geometric figure is a rectangle, and the diagonals of the rectangle clearly bisect each other visually.</perception> So yes, OC is equal to OD, which is 5 cm. So the answer is C. Then, let me check the options. The options are A.3, B.4, C.5, D.6, E. So the answer is C. So boxed C. Wait, but wait, maybe I'm missing something. <perception>The diagram clearly marks the length of OD as 5 cm</perception> The problem says "the length of OC". Since O is the midpoint, OC is half of diagonal BD. BD is $2 * OC$, but OD is 5 cm, so OC is 5 cm. That's correct. So the answer is C.

</think>

To solve the problem, we recall that in a rectangle, the diagonals are equal and bisect each other. This means the diagonals AC and BD intersect at point O, making O the midpoint of both diagonals. Thus, $OC = OD$. <perception>From the diagram, the segment OD is labeled 5 cm</perception>. Since O is the midpoint, OC is equal to OD. Therefore, the length of OC is 5 cm.

\boxed{C} (CORRECT)

References

- [1] Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025. [1](#)
- [2] Shuang Chen, Yue Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Haoxi Li, Manyuan Zhang, Jiayu Chen, Song Guo, and Nanyun Peng. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. *arXiv preprint arXiv:2510.08457*, 2025. [1](#)
- [3] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. [2](#)
- [4] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. [1](#)
- [5] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025. [1](#)
- [6] Shenzhi Wang Zhangchi Feng Dongdong Kuang Yuwen Xiong Yaowei Zheng, Junting Lu. Easyr1: An efficient, scalable, multi-modality rl training framework. <https://github.com/hiyouga/EasyR1>, 2025. [1](#)
- [7] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. [1](#)
- [8] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. [1](#)