

LongPerceptualThoughts: Distilling System-2 Reasoning for System-1 Perception

Yuan-Hong Liao[♣]Sven Elflein^{♣◇}Liu He[♣]Laura Leal-Taixé[◇]Yejin Choi[◇]Sanja Fidler^{♣◇}David Acuna[◇]

Abstract

Recent reasoning models through test-time scaling have demonstrated that long chain-of-thoughts can unlock substantial performance boosts in hard reasoning tasks such as math and code. However, the benefit of such long thoughts for system-2 reasoning is relatively less explored in other domains such as perceptual tasks where shallower, system-1 reasoning seems sufficient. In this paper, we introduce *LongPerceptualThoughts*, a new synthetic dataset with 30K long-thought traces for perceptual tasks. The key challenges in synthesizing elaborate reasoning thoughts for perceptual tasks are that off-the-shelf models are not yet equipped with such thinking behavior and that it is not straightforward to build a reliable process verifier for perceptual tasks. Thus, we propose a novel three-stage data synthesis framework that first synthesizes verifiable multiple-choice questions from dense image descriptions, then extracts simple CoTs from VLMs for those verifiable problems, and finally expands those simple thoughts to elaborate long thoughts via frontier reasoning models. In controlled experiments with a strong instruction-tuned 7B model, we demonstrate notable improvements over existing visual reasoning data-generation methods. Our model, trained on the generated dataset, achieves an average +3.4 points improvement over 5 vision-centric benchmarks, including +11.8 points on V* Bench. Notably, despite being tuned for vision tasks, it also improves performance on the text reasoning benchmark, MMLU-Pro, by +2 points.¹

1 Introduction

Reasoning models, such as OpenAI’s o1 (OpenAI et al., 2024) and Deepseek’s R1 (DeepSeek-AI et al., 2025), have demonstrated remarkable capabilities in solving complex reasoning problems by scaling test-time compute. Intuitively, they increase the number of tokens generated at inference-time, allowing the model to “think longer”—producing longer chain-of-thoughts (CoTs) that go beyond typical linear rationales that mimic textbook responses.

However, despite numerous attempts to match the performance of models like o1 or R1 on challenging math benchmarks (MAA, 2024; Lightman et al., 2023), less effort has been directed toward tasks beyond mathematical reasoning. Most notably, how to generate synthetic, long-form CoT reasoning traces that solve *vision-centric tasks*—suitable for both supervised fine-tuning and preference-based RL—remains largely underexplored.

Vision-centric tasks have proven challenging for vision-language models (VLMs), especially when the tasks require object counting and localization, scene understanding, and 2D/3D spatial reasoning (Liao et al., 2024a; Rahmzadehgervi et al., 2024; Campbell et al., 2024). Prior works have addressed these challenges by helping VLMs “see” better. Common approaches include modifying the input image (e.g., through cropping) or incorporating intermediate representations, into the CoT (Wu & Xie, 2024; Shao et al., 2024a; Wu et al.,

¹Project website: <https://andrewliao11.github.io/LongPerceptualThoughts>

[♣]University of Toronto, Vector Institute, [◇]NVIDIA, [♣]Purdue University

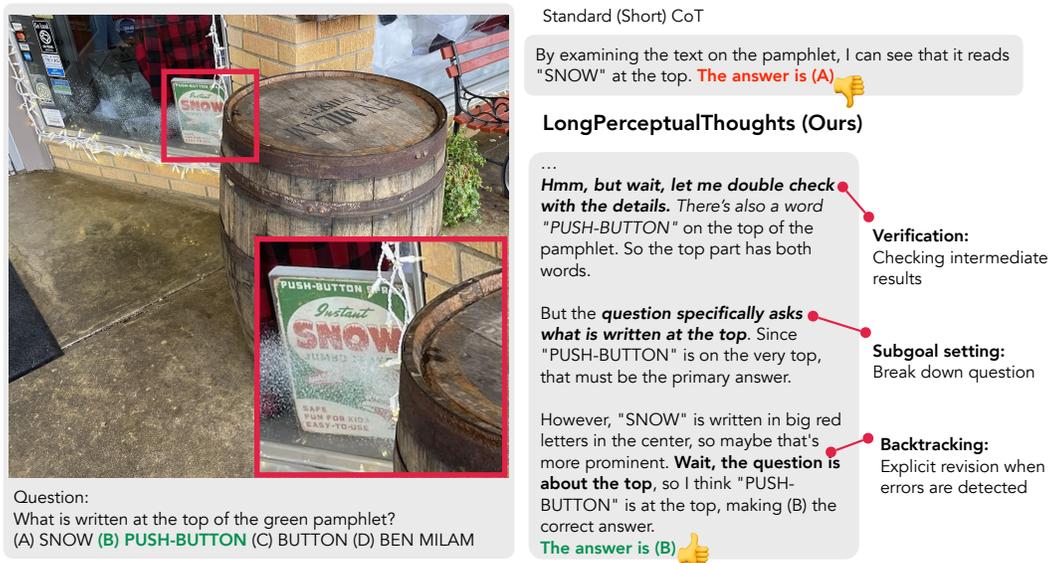


Figure 1: **LongPerceptualThoughts** is a new synthetic dataset with 30K long-thought traces for vision-centric tasks. Each trace contains diverse cognitive behaviors (e.g., verification, subgoal setting, and backtracking), akin to system-2 reasoning. CoTs generated by open-source VLMs often produce linear, rigid reasoning traces (top). In contrast, our novel data synthesis framework effectively expands these simple thoughts using frontier reasoning models, equipping VLMs with complex reasoning structures and rich cognitive behaviors—effectively distilling system-2 reasoning into instruction-tuned VLMs.

2025b). In contrast, we propose to synthesize data that implicitly equips VLMs with an internal search mechanism—one that unfolds through a textual inner monologue, enabling the model to explore multiple potential solution paths: revisiting different image regions, verifying intermediate conclusions, identifying inconsistencies, and self-correcting when necessary. Our approach is complementary to prior methods and mirrors the behavior observed in reasoning models like R1 and commercial VLMs such as o1, where reasoning performance improves by scaling test-time inference. We emphasize we do not claim that long textual CoTs are inherently superior or the only way to scale test-time inference in VLMs. Rather, our goal is to synthesize data to equip models with such capability—an approach shown to be effective in SoTA reasoning models. Furthermore, given the difficulty of building reliable process verifiers for search in perceptual tasks, our data-centric method offers a practical alternative.

In this work, we take a first step toward a scalable method that synthesizes long CoT data for vision-centric tasks. Specifically, we propose a novel three-stage data synthesis framework that: (1) generates *synthetic verifiable* multiple-choice questions from dense image captions, (2) extracts simple CoTs from VLMs for those questions, and (3) expands these simple CoTs into richer, long-form reasoning traces using frontier reasoning models. Notably, our framework performs *three layers of synthesis*: one to generate questions, one to think, and the last one to think harder. As shown in Fig. 1, using our framework, we generate **LongPerceptualThoughts**, a dataset of 30k examples for both supervised fine-tuning (SFT) and direct preference optimization (DPO), and use it to fine-tune a strong instruction-tuned VLM. The resulting model shows an average +3.4 points improvement across 5 vision-centric benchmarks, including a gain of +11.8 points on V* Bench, while typical multimodal reasoning datasets fail to improve the base VLM due to overthinking. Notably, despite being tuned for vision tasks, it also improves on the challenging text reasoning benchmark MMLU-Pro by +2 points.

2 Synthesize Long CoT Data for Vision-Centric Tasks

In this section, we introduce a novel data synthesis framework to synthesize long chain-of-thought (CoT) data for fine-tuning a vision-language model (VLM). Inspired by DeepSeek’s R1, we are interested in collecting data consisting of thoughts and answers in the format of `<think> thought </think> <answer> answer </answer>`. We start by discussing two desired properties of reasoning data for vision-centric tasks in Sec. 2.1. Based on these two properties, we explain our data synthesis framework in Sec. 2.3. Finally, we use the synthesized long CoT data to construct LongPerceptualThoughts, consisting of both SFT and preference datasets in Sec. 2.4.

2.1 Desired Properties in Long Chain-of-Thought

Inspired by the recent success in OpenAI’s o1 and DeepSeek’s R1, we further define *Long CoT* as an extended, structured rationale that mirrors how a human might approach complex visual reasoning tasks. Unlike the short, linear responses typically produced by current open-source VLMs, Long CoTs explore alternative solutions, verifying intermediate steps, and adjusting course when necessary. Drawing on the framework proposed in Gandhi et al. (2025), we characterize Long CoTs in vision-centric tasks through three core cognitive behaviors: **verification** (checking intermediate conclusions for correctness), **backtracking** (recognizing and revising failed solution paths), and **subgoal setting** (breaking down the task into smaller, solvable components). These cognitive behaviors have been observed in LLM to increase performance by scaling test-time compute (Muennighoff et al., 2025).

To study cognitive behaviors in vision-centric tasks, we begin by analyzing the outputs of strong instruction-tuned VLMs, following Gandhi et al. (2025). Despite its general capabilities, the model rarely displays the cognitive behaviors described earlier. The responses are often shallow and rigid, lacking the iterative, self-corrective reasoning we aim to capture. Figure 3a quantifies this gap between the response from Qwen2.5-VL-7B-Instruct and Gemini 2.0 Flash Thinking. At the end of this section, we introduce LongPerceptualThoughts that drastically diversifies the standard CoT with the desired cognitive behaviors.

2.2 Preliminaries

Formally, given an **image** v , our goal is to construct a quadruple (v, q, Z, a) , consisting of a **question** q , a long CoT **reasoning trace** Z , and a **final answer** a . We also assume the access to dense image **descriptions** c . A long CoT is composed of multiple thoughts that incorporate cognitive behaviors such as backtracking, verification, and subgoal setting. Formally, we define a long CoT as a sequence of intermediate thoughts: $Z := z_1 \oplus z_2 \oplus \dots$, where \oplus denotes concatenation and z is a sequence of sentences, typically delineated by double new lines, *i.e.* “`\n\n`”.

For preference data—used in reinforcement learning—our goal is to construct a preference pair of $(v, q, Z^+, a^+) \succ (v, q, Z^-, a^-)$, where the superscripts $+$ and $-$ indicate the preferred (correct) and non-preferred (incorrect or suboptimal) reasoning trajectories and their answers, and \succ denotes that the left-hand tuple is preferred over the right-hand one.

2.3 Thought-Expansion: Distilling System-2 Reasoning into Instruction-Tuned VLMs

For an image v , we begin by assuming access to its dense image description c that provides comprehensive visual features in the image. While in principle, one could also obtain such descriptions using a captioning model, here we assume access to such a dataset *e.g.*, DOCCI (Onoe et al., 2024) or DCI (Urbanek et al., 2023). In our proposed data synthesis framework, we leverage three foundation models: an LLM, a VLM that takes interleaved image and text as input and generates text, and a reasoning LLM that explicitly produces thoughts and answers. We use \mathcal{M}_{LLM} , \mathcal{M}_{VLM} , and $\mathcal{M}_{\text{Reason}}$ to denote them, respectively.

Below, we describe the three key stages of our data synthesis process.

Stage 1: Convert dense descriptions to multiple-choice questions We first convert dense descriptions into multiple-choice questions (MCQs) using an LLM. Specifically, we prompt

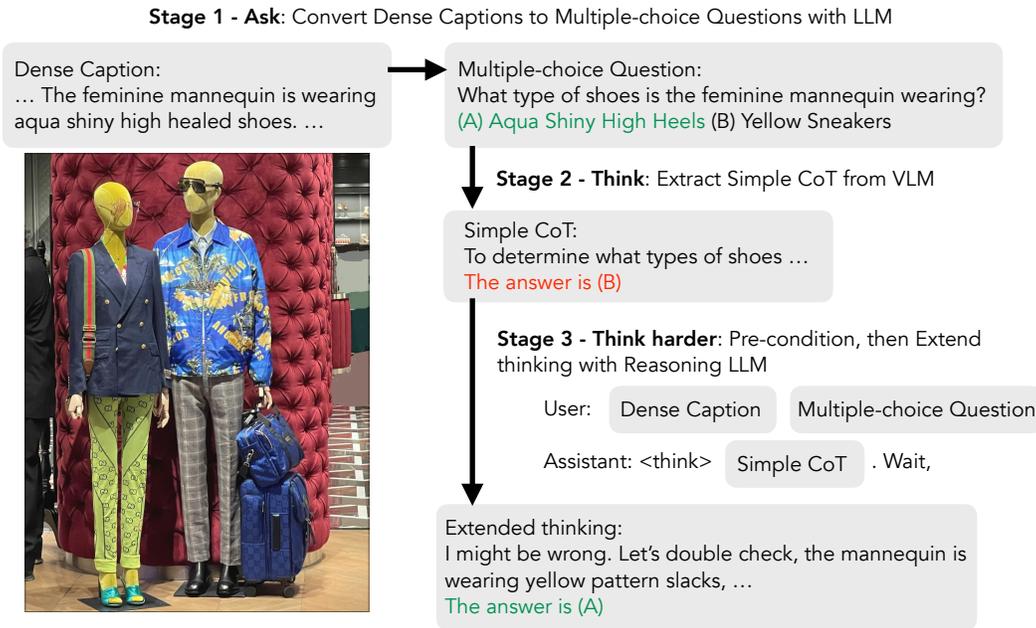


Figure 2: **Ask, Think, and Think Harder: The three stages to synthesize long CoT data for vision-centric tasks.** Assuming the access to an image and its associated dense caption, we first ask an LLM to convert dense captions to multiple-choice questions. In Stage 2, we extract simple CoT from VLM. These simple CoTs typically exhibits shallow and rigid reasoning, especially in vision-centric tasks. Therefore, in Stage 3, we precondition a reasoning LLM with these simple CoTs and append a subtle cue, *e.g.*, “Wait,” to elicit more diverse long CoTs.

\mathcal{M}_{LLM} to generate MCQs based on an image and its associated dense descriptions. This step offers two key advantages that are leveraged in subsequent stages: (1) It ensures that each generated question is answerable using only the dense descriptions, allowing us to synthesize the reasoning process purely from the text modality. (2) The multiple-choice format enables easy identification of prediction correctness, which is essential for constructing positive and negative pairs in our preference dataset. Formally, this step produces a triplet $(v, q, a^*) := \mathcal{M}_{LLM}(v, c)$. We use gpt-4o-mini as \mathcal{M}_{LLM} to balance the cost and the quality of MCQs.

Stage 2: Extract Simple CoTs from VLM To generate long CoTs that the VLM is familiar with, we use the same VLM that will later be fine-tuned. Specifically, we prompt \mathcal{M}_{VLM} with the image and the multiple-choice question generated in Stage 1 to produce a rationale and a final prediction, denoted as $(z_1, a_1) := \mathcal{M}_{VLM}(v, q)$. Sampling from the same VLM ensures that the synthesized CoTs remain within the model’s output distribution, which we observed to be a key factor in downstream performance. By comparing the predicted answer a_1 with the ground-truth answer a^* from Stage 1, we can further categorize the data into positive (z_1^+, a_1^+) or negative examples (z_1^-, a_1^-) . These can then be reused to construct either a SFT or a preference dataset. This process is akin to the rejection sampling in self-training algorithms such as RFT (Yuan et al., 2023) and STaR (Zelikman et al., 2022). We choose Qwen2.5-VL-7B-Instruct as our \mathcal{M}_{VLM} , as the Qwen2.5 series has demonstrated a non-trivial probability of exhibiting cognitive behaviors (Gandhi et al., 2025).

Stage 3: Thought-Expansion using the Reasoning Model. The analysis in Fig. 3a reveals that CoTs sampled from open-source VLMs typically exhibit shallow and rigid reasoning, with limited exploration in the output space. Inspired by the diverse cognitive behaviors observed in the responses of frontier reasoning models, we aim to leverage a reasoning LLM to generate long CoTs. However, naively sampling from \mathcal{M}_{Reason} can produce CoTs that deviate significantly from the output distribution in VLM, which may degrade the performance of instruction-tuned models during fine-tuning. The similar findings have

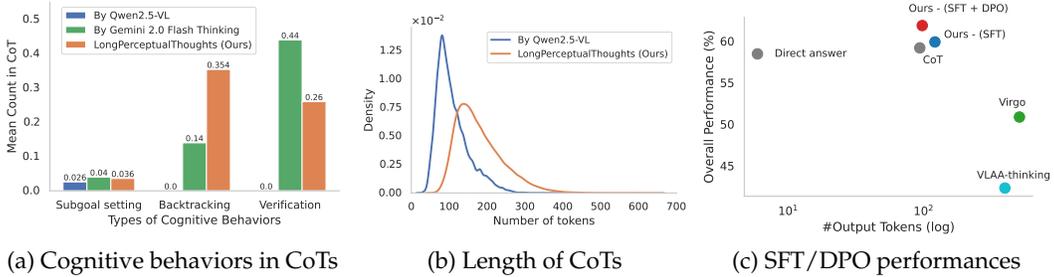


Figure 3: **(a) Analysis of Cognitive Behaviors in Chain-of-Thought (CoT).** CoTs from open-source VLMs often follow rigid structures. In contrast, frontier reasoning VLMs—such as Gemini 2.0 Flash Thinking—exhibit more diverse cognitive behaviors, including subgoal setting, backtracking, and verification. Our introduced long CoT dataset, LongPerceptualThoughts, also demonstrates a wide range of such behaviors. **(b) Length of CoTs.** The CoTs in LongPerceptualThoughts are significantly longer than those generated by popular VLMs, e.g. Qwen2.5-VL. **(c) Response length vs. aggregated performances.** Fine-tuning VLM on LongPerceptualThoughts with complex reasoning structures lead to higher overall performances with slightly more output tokens. On the other hand, fine-tuning on other multimodal reasoning leads to over-thinking and worse performance. Cognitive behaviors are quantified following Gandhi et al. (2025).

been discovered in LLMs as well (Ren et al., 2024; Li et al., 2025; Wu et al., 2025a). To address this, we introduce a *thought-expansion* mechanism that guides the reasoning LLM $\mathcal{M}_{\text{Reason}}$ to extend the CoT produced in Stage 2, while injecting cognitive behaviors such as backtracking, verification, and subgoal setting. Specifically, we precondition $\mathcal{M}_{\text{Reason}}$ with the CoT generated by VLM z_1 and append a subtle cue—selected from a set of predefined markers m (e.g., “Wait,” “Hmm,” “Alternatively,”)—to elicit more reflective or exploratory responses. Formally, we structure the prompt as:

$$\begin{aligned} \text{User: } & c \oplus q \\ \text{Assistant: } & \langle \text{think} \rangle \oplus z_1 \oplus m \end{aligned}$$

and ask $\mathcal{M}_{\text{Reason}}$ to continue the thought to obtain (z_2, a_2) . This approach enables the reasoning LLM to expand the familiar reasoning traces while enriching them with non-linear problem-solving behaviors. Similar to Stage 2, we can also use a_2 to categorize the data into positive or negative examples. Fig. 2 demonstrates the way to construct such prompt visually. We use DeepSeek-R1-Distill-Qwen-32B as our $\mathcal{M}_{\text{Reason}}$, as it is derived from the same Qwen2.5 series as the \mathcal{M}_{VLM} . For more details, see the prompt template in Appendix F.7.

Our proposed framework is scalable and only assumes the access to dense image description datasets. From Stage 1 to Stage 3, for an image v and its associated descriptions c , we obtain MCQs (q, a^*) in Stage 1. Then, in Stage 2 and 3, we obtain two intermediate thoughts and their associated predicted answers (z_1, z_2, a_1, a_2) . Finally, we have long CoT data Z obtained by: $z_1 \oplus m \oplus z_2$. We will omit m in the following sections for the sake of brevity and clarity.

2.4 Construct SFT and DPO Datasets

In Sec. 2.3, we described a process to obtain long-form CoTs that not only aligns with the VLM’s output distribution but also contains system-2 reasoning behaviors. To construct a supervised fine-tuning (SFT) dataset, we collect CoTs that lead to correct predictions. This includes examples of the form:

$$(z_1^+, a_1^+), (z_1^+ \oplus z_2^+, a_2^+), (z_1^- \oplus z_2^+, a_2^+)$$

To construct a preference dataset, we follow Setlur et al. (2024); Zhang et al. (2025) and define pairwise preferences based on correctness and compactness Team et al. (2025). Specifically:

Correctness:

$$\begin{aligned} (z_1^+, a_1^+) &\succ (z_1^-, a_1^-) \\ (z_1^- \oplus z_2^+, a_2^+) &\succ (z_1^-, a_1^-) \end{aligned}$$

Compactness:

$$(z_1^+, a_1^+) \succ (z_1^+ \oplus z_2^+, a_2^+)$$

Akin to [Setlur et al. \(2024\)](#), by constructing the preference pairs of $(z_1^- \oplus z_2^+, a_2^+) \succ (z_1^-, a_1^-)$, we encourage the model to increase the likelihood $P(z_2^+, a_2^+ | z_1^-)$ and decrease the likelihood $P(a_1^- | z_1^-)$, leading to better credit assignment.

Filtering. Since z_2 is generated by a reasoning LLM using dense captions as input, it may include phrases like “As the description says.” To address this, we define a list of “bad words” and filter out any thoughts containing them.

Details of LongPerceptualThoughts. We use 500 images and their dense captions from DOCCI. Stage 1 produces 4590 multiple-choice questions (MCQs). For long CoT data, we construct an SFT dataset with 30295 examples and a preference dataset with 17208 pairs, following filtering and deduplication. We use gpt-4o-mini as \mathcal{M}_{LLM} , Qwen2.5-VL-7B-Instruct as \mathcal{M}_{VLM} , and R1-Distill-Qwen-32B as \mathcal{M}_{Reason} .

3 Experiments

In this section, we first describe the experimental setup on five vision-centric benchmarks in Sec. 3.1 and present our main results in Sec. 3.2. In Sec. 3.3, we go beyond vision by evaluating our fine-tuned VLMs on a challenging text-only benchmark. Lastly, in Sec. 3.4, we analyze the response of the fine-tuned VLMs.

3.1 Setup

Model. We use Qwen2.5-VL-7B-Instruct ([Bai et al., 2025](#)) as our base model to fine-tune throughout the paper. For the sake of brevity, we refer to it as BaseModel in this section. We adopt full-parameter fine-tuning using LLaMA-factory ([Zheng et al., 2024](#)). See more training details in Appendix F.

Benchmarks. We evaluate our models on vision-centric tasks. For benchmarks covering general knowledge, we only keep their vision-centric splits, such as MME-RealWorld ([Zhang et al., 2024](#)) and MMStar ([Onoe et al., 2024](#)). To better clarify the differences, we refer to them as MME-RW-V and MMStar-V, respectively. Additionally, following [Tong et al. \(2024a\)](#), we include the vision-centric benchmarks: CV-bench, V* Bench, and MMVP, that involve 2D/3D spatial reasoning, fine-trained attribution, coarse scene understanding, *etc.* In total, the benchmarks consist of 10284 images and 15315 questions. More details are in Appendix B.

Evaluation metrics. All the benchmarks used in this work are in multiple-choice question format. We standardize their format and use regex to extract the answers. We report accuracy across all benchmarks.

Baselines. To explore the vision-centric capabilities of BaseModel, we evaluate its zero-shot predictions and apply a prompt-based chain-of-thought approach. Specifically, we prompt the model to generate `<think> thought </think>` before producing an answer—a method we refer to as Internal Thinking CoT.

For multimodal datasets, we compare LongPerceptualThoughts with one captioning dataset, DOCCI, and two multimodal reasoning datasets, Virgo ([Du et al., 2025](#)) and VAAL-thinking ([Chen et al., 2025](#)). (1) DOCCI is a human-annotated dense caption dataset, highlighting comprehensive descriptions for images. For a fair comparison with LongPerceptualThoughts, we use the exact same set of 500 images and their captions as training data. (2) Virgo distills reasoning capabilities from the language model QwQ ([Team, 2024b](#)) and the multimodal model QvQ ([Team, 2024a](#)). We adopt Virgo’s self-distillation split, generated

| Approach | Avg | CV-Bench | V* Bench | MMVP | MMStar-V | MME-RW-V |
|--|--------------|--------------|----------------|---------------|---------------|---------------|
| Qwen2.5-VL-7B-Instruct | 58.47 | 74.74 | 48.51 | 73.67 | 63.73 | 31.68 |
| + Internal Thinking CoT | 59.18 | 75.42 | 55.08 | 70.60 | 62.40 | 32.40 |
| + DOCCI | 36.14 | 50.82 | 39.96 | 48.67 | 8.67 | 32.58 |
| + VLAA-thinking | 42.32 | 68.50 | 53.53 | 66.67 | 0.53 | 22.38 |
| + Virgo | 50.87 | 67.22 | 44.14 | 57.67 | 57.60 | 27.71 |
| + LongPerceptualThoughts- SFT (Ours) | 59.90 | 76.05 | 60.53 (+12.02) | 70.00 | 60.67 | 32.25 |
| + LongPerceptualThoughts- SFT + DPO (Ours) | 61.87 (+3.4) | 76.61 (+1.8) | 60.31 (+11.8) | 75.00 (+1.33) | 64.00 (+0.27) | 33.45 (+1.77) |

Table 1: **Main results on five vision-centric benchmarks.** We group the approaches into three categories: training-free methods, existing multimodal reasoning datasets, and our proposed LongPerceptualThoughts. On vision-centric tasks, fine-tuning on other multimodal reasoning datasets often leads to poorer performance, likely due to reduced instruction-following ability, domain mismatch, or an inability to capture the complex reasoning learned by larger models. In contrast, fine-tuning on LongPerceptualThoughts yields an average improvement of +1.5 points, and this gain increases to +3.4 points when using preference pairs. Notably, it achieves a 12-point improvement on the challenging V* Bench.

by first distilling QwQ into Qwen2-VL-72B-Instruct, then using the fine-tuned model to produce reasoning data for multimodal questions. The Virgo dataset includes 14, 540 examples. (3) VLAA-thinking generates multimodal reasoning data by prompting DeepSeek’s R1 model with additional caption information. It contains 158k examples, from which we randomly sample 25k for training to match a similar size to our dataset. ²

3.2 Main Results

We report aggregated performances in Table 1 and detail our main findings on five vision-centric benchmarks:

LongPerceptualThoughts consistently improves performance on vision-centric benchmarks by +3.4 points via DPO. We first perform supervised fine-tuning on the synthesized LongPerceptualThoughts. Across 5 benchmarks, we improve BaseModel by nearly +1.5 points on average, especially in challenging tasks such as V* bench, improving by +12 points. However, the improvements on the rest of the benchmarks are marginal. We hypothesize that this is due to noisy or erroneous tokens in our SFT datasets, which may hurt fine-tuning performance. While several prior works suggest the impacts of such erroneous tokens are marginal, they either work on models under 300M parameters (Ye et al., 2024) or target different aspects, such as safety alignment (Zhang et al., 2025). In this work, we try not to over-engineer the training recipe to highlight the effectiveness of the synthesized datasets. Unlike VLAA-thinking and Virgo that simply distill knowledge from reasoning LLMs or VLMs, our data generation pipeline allows us to construct preference data. By fine-tuning on these preference pairs, the aforementioned erroneous tokens might naturally be mitigated. For example, by performing preference-based fine-tuning such as DPO, on $(z_1^- \oplus z_2^+, a_2^+) \succ (z_1^-, a_1^-)$, the model should naturally increase the likelihood of $P(z_2^+, a_2^+ | z_1^-)$ as opposed to $P(a_1^- | z_1^-)$. This helps the model reduce the impact of erroneous tokens. We find that by first applying SFT and then DPO, we obtain consistent improvements across all evaluation datasets, improving by +3.4 accuracy points. Overall, we find that LongPerceptualThoughts generally leads to consistent improvements and the preference data is the key to bring up the improvements.

Off-the-shelf captioning data hurts instruction-tuned VLMs on vision-centric benchmarks. Since LongPerceptualThoughts is derived from DOCCI, we are interested to see if fine-tuning BaseModel on DOCCI improves. Table 1 shows that training on DOCCI results in inferior performances. Perhaps expected, we find that fine-tuning on DOCCI alone especially leads to bad instruction following.

Off-the-shelf distillation hurts performance on vision-centric benchmarks. Both Virgo and VLAA-thinking are multimodal reasoning datasets. VLAA-thinking is distilled from

²We accessed the dataset in mid-March 2025.

R1 with the help of image captions. Virgo is distilled from a fine-tuned VLM distilled from QwQ. While both datasets are equipped with complex reasoning structures, finetuning the BaseModel does not improve vision-centric performance; instead, finetuning on VLAA-thinking and Virgo hurts the performances by -16.15 and -7.6 points, respectively. We hypothesize that there are two reasons that lead to the performance drops: (1) Both datasets are distilled from a much much larger LLMs (671B R1 model) or VLMs (Qwen2-VL-72B-Instruct), potentially resulting in the learnability gap (Li et al., 2025). (2) In particular, the multimodal reasoning data from Virgo is math-focused. We hypothesize that there is a gap between reasoning over perceptual tasks and math-related tasks. On the other hand, VLAA-thinking consists of a diverse set of datasets including DocVQA (Mathew et al., 2021), ChartVQA (Masry et al., 2022), etc. When using reasoning data exclusively from more natural image sources, we surprisingly observe worse performance than random subsampling. See Appendix D for details.

3.3 Beyond Vision: Evaluation on the Text-Only Reasoning Benchmark

Following the same setup in Sec. 3.1, we evaluate VLMs fine-tuned on multimodal reasoning training datasets in out-of-distribution (OOD) tasks. In particular, we adopt MMLU-Pro, a challenging text-only reasoning benchmark.

MMLU-Pro (Wang et al., 2024a). MMLU-Pro is built on top of MMLU (Hendrycks et al., 2021) by integrating more reasoning-focused questions and expanding the choices set. MMLU-Pro spans 14 diverse domains including mathematics, physics, chemistry, etc., encompassing over 12000 questions.

Results. As shown in Table. 2, we find that BaseModel fine-tuned on LongPerceptualThoughts surprisingly improves on these text-only reasoning tasks, with an average gain of +2 points. In contrast, VLAA-thinking and Virgo hurt performance, suggesting that directly distilling from stronger teachers may lead to sharp drops in OOD tasks. We propose two hypotheses for LongPerceptualThoughts’ effectiveness: (1) it introduces complex reasoning structures that improve BaseModel’s general reasoning abilities; and (2) it remains close to the original output distribution, making the new reasoning skills easier to learn without disrupting existing knowledge. Additional MMLU-Pro evaluation details are provided in Appendix E.

| Approach | Acc |
|---------------------------|--------------|
| Qwen2.5-VL-7B-Instruct | - |
| + CoT | 48.07 |
| + DOCCI | 32.99 |
| + VLAA-thinking | 21.56 |
| + Virgo | 37.95 |
| + Ours - SFT | 50.77 |
| + Ours - SFT + DPO | 50.20 |

Table 2: Evaluation on out-of-distribution tasks text-only reasoning benchmark MMLU-Pro.

3.4 Analysis on Fine-tuned VLM Responses

To better understand our fine-tuned VLM, we visualize its responses with two key factors: aggregated performances and question difficulties.

Response length vs. performances. There has been growing interest in how LLMs leverage test-time compute. To investigate this, we aggregated response lengths and performance across five vision-centric benchmarks. Fig. 3c illustrates the relationship between test-time compute—measured by response length—and model performance. We observe that VLMs fine-tuned on LongPerceptualThoughts tend to produce slightly longer responses, especially after SFT. Interestingly, DPO training results in shorter responses, which aligns with the compactness encouraged during DPO pair construction. One possible direction is to exclude such preference pairs to allow models to make fuller use of test-time compute. In contrast, other multimodal reasoning benchmarks reveal signs of overthinking, where models generate unnecessarily lengthy responses.

Response length vs. question difficulty. Another desirable characteristic of the thinking process in LLMs is their ability to allocate appropriate “thinking time” based on a question’s difficulty. Following prior works (Lightman et al., 2024; Snell et al., 2025), we define question difficulty with respect to a base VLM, i.e., Qwen2.5-VL-7B-Instruct. For each question, we estimate the model’s accuracy using 11 samples and bin the questions into two quantiles:

easy and hard. Our analysis focuses on the outputs of the VLM fine-tuned via DPO on LongPerceptualThoughts. We observe that the model naturally allocates more test-time compute—reflected in longer responses—for harder questions, where its original (pre-fine-tuning) performance was worse. See Appendix C for details and visualization.

4 Related Work

Reasoning in Large Language Models. The complex reasoning abilities of large language models (LLMs) have been uncovered through various approaches. Chain-of-thought (CoT) prompting elicits their intrinsic reasoning capabilities, improving performance on language-based causal reasoning tasks (Wei et al., 2022), and has been extended into tree-based searches to enhance reasoning further (Yao et al., 2023). Similar search-like behavior can be induced through verifier guidance (Lifshitz et al., 2025), curated datasets (Shao et al., 2024b), or supervised fine-tuning on synthetic reasoning data (Gandhi et al., 2024; Lehnert et al., 2024). More recently, DeepSeek-R1 (Guo et al., 2025) achieved state-of-the-art reasoning through reinforcement learning, exhibiting human-like traits such as self-correction and verification. In contrast, s1 (Muennighoff et al., 2025) improves mathematical reasoning via supervised fine-tuning on 1000 distilled reasoning traces from frontier models. While most prior work focuses on math and coding tasks, our goal is to explore how such strong reasoning capabilities can be effectively applied to perception.

Reasoning in Vision-Centric Tasks. Unlike reasoning in math or coding tasks, vision-centric problems often involve significant uncertainty due to partial information, perceptual noise, and visual ambiguities. Prior works primarily address this by helping VLMs “see” better. For instance, SEAL (Wu & Xie, 2024) uses a search-like cropping mechanism to iteratively navigate an image, while VisualCoT (Shao et al., 2024a) generates auxiliary visual cues to guide attention. Other approaches (Wang et al., 2024b; Liao et al., 2024b) decompose complex tasks into simpler verification steps to enhance model robustness. In contrast, we aim to teach VLMs to reason better—encouraging them to explore multiple solution paths by revisiting image regions, verifying intermediate conclusions, and engaging in textual inner monologue. Concurrent work on multimodal reasoning addresses this challenge, particularly in math problem solving, using techniques such as distillation from advanced reasoning LLMs (Du et al., 2025; Thawakar et al., 2025) and reinforcement learning (Liu et al., 2025; Huang et al., 2025). In this work, we study how system-2 reasoning can improve vision-centric tasks, and propose a data synthesis framework that generates long CoT examples to teach visual reasoning through deliberate, step-by-step thinking in the textual space.

5 Conclusions

In this work, we explore how system-2 reasoning can enhance vision-centric tasks. We introduce a novel, scalable data synthesis framework that requires only dense image captions. The framework generates verifiable multiple-choice questions, extracts simple chains of thought (CoTs) from vision-language models (VLMs), and expands them into rich, long-form reasoning traces using frontier reasoning models. This process yields LongPerceptualThoughts, a synthetic dataset containing 30k detailed reasoning traces for perceptual tasks. Fine-tuning Qwen2.5-VL-7B-Instruct on LongPerceptualThoughts improves performance by +3.4 points across five vision benchmarks, including an +11.8-point gain on V* Bench. Remarkably, despite being trained on vision tasks, the model also improves by +2 points on the out-of-distribution text-only reasoning benchmark MMLU-Pro.

Acknowledgements

We thank Rafid Mahmood, Jaehun Jung, Jen-Hao Cheng, Ali Hatamizadeh, Ximing Lu, Hyunwoo Kim and Amlan Kar for their helpful comments and feedback on an early discussions and draft of this paper.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=Q5RYn6jagC>.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025. URL <https://arxiv.org/abs/2504.11468>.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanji Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhenwen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm, 2025. URL <https://arxiv.org/abs/2501.01904>.
- Kanishk Gandhi, Denise H J Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah Goodman. Stream of search (sos): Learning to search in language. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=2cop2jmQVL>.

- Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars, 2025. URL <https://arxiv.org/abs/2503.01307>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL <https://arxiv.org/abs/2503.06749>.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Lucas Lehnert, Sainbayar Sukhbaatar, Dijia Su, Qinqing Zheng, Paul McVay, Michael Rabbat, and Yuandong Tian. Beyond a*: Better planning with transformers via search dynamics bootstrapping. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=SGoVIC0u0f>.
- Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners, 2025. URL <https://arxiv.org/abs/2502.12143>.
- Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17028–17047, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.947. URL <https://aclanthology.org/2024.emnlp-main.947/>.
- Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Can feedback enhance semantic grounding in large vision-language models?, 2024b. URL <https://arxiv.org/abs/2404.06510>.
- Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers, 2025. URL <https://arxiv.org/abs/2502.20379>.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=v8L0pN6EOi>.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME 2024*, February 2024. URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>.

- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL <https://aclanthology.org/2022.findings-acl.177/>.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *ECCV*, 2024.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng,

- Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024.
- Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=4XIKfvNYvx>.
- Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, December 2024.
- Xuan Ren, Biao Wu, and Lingqiao Liu. I learn better if you speak my language: Understanding the superior performance of fine-tuning large language models with LLM-generated responses. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10225–10245, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.571. URL <https://aclanthology.org/2024.emnlp-main.571/>.
- Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. RL on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold, 2024. URL <https://arxiv.org/abs/2406.14532>.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0ff38d72a2e0aa6dbe42de83a17b2223-Paper-Datasets_and_Benchmarks_Track.pdf.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024b. URL <https://arxiv.org/abs/2402.03300>.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=4FWAwZtd2n>.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL <https://arxiv.org/abs/2501.12599>.

- Qwen Team. Qvq: To see the world with wisdom, December 2024a. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.
- Qwen Team. Qvq: Reflect deeply on the boundaries of the unknown, November 2024b. URL <https://qwenlm.github.io/blog/qvq-32b-preview/>.
- Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-01: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025. URL <https://arxiv.org/abs/2501.06186>.
- Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL <https://openreview.net/forum?id=Vi8AepAXGy>.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, June 2024b.
- Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26690–26699, 2023. URL <https://api.semanticscholar.org/CorpusID:266209761>.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL <https://openreview.net/forum?id=y10DM6R2r3>.
- Yuxuan Wang, Alan Yuille, Zhuowan Li, and Zilong Zheng. Exovip: Step-by-step verification and exploration with exoskeleton modules for compositional visual reasoning. In *First Conference on Language Modeling*, 2024b. URL <https://openreview.net/forum?id=6U1FEKP7Ar>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- Chao-Chung Wu, Zhi Rui Tam, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. Clear minds think alike: What makes llm fine-tuning robust? a study of token perplexity, 2025a. URL <https://arxiv.org/abs/2501.14315>.
- Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, June 2024.
- Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models, 2025b. URL <https://arxiv.org/abs/2503.12799>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.),

- Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.
- Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems, 2024. URL <https://arxiv.org/abs/2408.16293>.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.
- Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.
- Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel M. Bikel, Jason E Weston, and Eric Michael Smith. Backtracking improves generation safety. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=Bo62NeU6VF>.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.38. URL <https://aclanthology.org/2024.acl-demos.38/>.

| | CV Bench | V* Bench | MMVP | MMStar-V | MME-RealWorld-V | Total |
|-------------|----------|----------|------|----------|-----------------|-------|
| # Images | 2638 | 191 | 300 | 750 | 6405 | 10284 |
| # Questions | 2638 | 191 | 300 | 750 | 11436 | 15315 |

Table 3: Vision-centric benchmark statistics.

A Table of Content

1. Sec. B elaborates the details of the considered five vision-centric benchmarks
2. Sec. D provides additional experimental results including additional comparison with self-training and our efforts to improve VLAA-thinking and Virgo.
3. Sec. E provides the full evaluation results on text-only reasoning benchmark, MMLU-Pro.
4. Sec. F provides implementation details in dataset generation, VLM training, and VLM inference.
5. Sec. G provides additional qualitative results of our dataset generation pipeline.
6. Sec. C provides the analysis of fine-tuned VLM’s response length versus question difficulties.

B Benchmark details

We describe the details of each benchmark:

1. CV-Bench (Tong et al., 2024a) is a comprehensive benchmark of over 2k manually-inspected examples, evaluating visual understanding across domains such as object recognition, scene understanding, and visual reasoning.
2. V* Bench (Wu & Xie, 2024) targets fine-grained visual reasoning tasks that demand detailed analysis of visual elements.
3. MMVP (Tong et al., 2024b) tests visual pattern recognition using “CLIP-blind pairs”—visually distinct images perceived as similar by CLIP—highlighting systematic limitations in VLMs.
4. MMStar-V includes tasks assessing instance-level reasoning, fine-grained perception (detecting subtle visual details), and coarse perception (understanding overall scene context).
5. MME-RW-V. MME-RealWorld evaluates real-world visual understanding across domains such as autonomous driving, remote sensing, monitoring, diagrams, tables, and OCR, testing both perception and reasoning. From these, we select three perception tasks—Remote Sensing, Monitoring, and Autonomous Driving—and two reasoning tasks—Monitoring and Autonomous Driving—to form MME-RealWorld-V.

As a result, our evaluation provides a comprehensive view on the perceptual capabilities enabled by the training datasets under consideration. Table 3 shows the basic statistics of the considered benchmarks.

C Response length vs. question difficulty

Following prior works , we define question difficulty with respect to a base VLM, *i.e.* Qwen2.5-VL-7B-Instruct. For each question, we estimate the model’s accuracy using 11 samples and bin the questions into two quantiles: easy and hard. Our analysis focuses on the outputs of the VLM fine-tuned via DPO on LongPerceptualThoughts. Fig 4 shows the distribution of response lengths across the easy and hard bins for four different tasks. We observe that the model naturally allocates more test-time compute—reflected in longer

| Approach | Avg | CV-Bench | V* Bench | MMVP | MMStar-V | MME-RW-V |
|------------------------|-------|----------|----------|-------|----------|----------|
| Qwen2.5-VL-7B-Instruct | 58.47 | 74.74 | 48.51 | 73.67 | 63.73 | 31.68 |
| VLAA-thinking | 42.32 | 68.50 | 53.53 | 66.67 | 0.53 | 22.38 |
| + only natural images | 34.96 | 61.91 | 28.86 | 55.00 | 6.20 | 22.86 |
| Virgo | 50.87 | 67.22 | 44.14 | 57.67 | 57.60 | 27.71 |
| + improved formatting | 52.58 | 68.94 | 46.54 | 66.33 | 53.47 | 27.60 |

Table 4: Attempted improvements on top of VLAA-Thinking and Virgo baselines.

| | Avg | Biology | Business | Chemistry | CompSci. | Econ. | Engin. | Health | History | Law | Math | Phil. | Physics | Psych. | Other |
|------------------------|-------|---------|----------|-----------|----------|-------|--------|--------|---------|-------|-------|-------|---------|--------|-------|
| Qwen2.5-VL-7B-Instruct | 48.07 | 68.62 | 55.77 | 44.79 | 49.51 | 61.26 | 34.26 | 47.68 | 43.57 | 24.89 | 50.41 | 38.88 | 47.19 | 60.65 | 45.56 |
| DOCCI | 32.99 | 51.60 | 42.33 | 22.61 | 37.32 | 43.48 | 18.89 | 32.76 | 22.31 | 10.26 | 40.19 | 29.46 | 25.56 | 51.13 | 33.98 |
| VLAA-Thinking | 21.56 | 25.24 | 27.76 | 15.11 | 20.73 | 25.47 | 7.64 | 24.45 | 29.40 | 13.35 | 26.72 | 20.04 | 17.78 | 21.43 | 26.73 |
| Virgo | 37.95 | 64.02 | 44.36 | 28.98 | 36.59 | 50.36 | 10.63 | 38.63 | 37.27 | 21.16 | 41.67 | 33.07 | 33.18 | 53.88 | 37.45 |
| Ours - SFT | 50.77 | 71.83 | 56.78 | 50.35 | 51.22 | 62.68 | 38.49 | 50.86 | 42.78 | 25.07 | 64.25 | 40.88 | 50.65 | 60.78 | 44.16 |
| Ours - SFT + DPO | 50.20 | 73.08 | 55.26 | 45.94 | 48.29 | 62.09 | 37.98 | 51.10 | 45.41 | 28.25 | 59.07 | 40.68 | 48.73 | 62.28 | 44.70 |

Table 5: Results for all categories of the MMLU-Pro dataset.

responses—for harder questions, where its original (pre-fine-tuning) performance was worse.

D Additional Results

VLAA-Thinking and Virgo adjustments. As we saw degradation in performance when training on both, Virgo and VLAA-Thinking, we spent additional time investigating the datasets and the model behavior they are causing which lead to these results.

We found that VLAA-Thinking consists of large proportions of math questions whereas natural image data is dominating the considered benchmarks as we focus on perceptual tasks. We hypothesize that this distribution shift might lead to lower performance. To investigate, we consider a version of VLAA-Thinking where we only keep the image subsets containing natural images, *i.e.*, ALLaVA-LAION and VizWiz, and randomly sample a subset of the same size. For Virgo, we found that predictions would not consistently respect the system prompt when formatting answers leading to inconsistencies with our regex-based evaluation. We thus explore a version of the dataset where we only copy the answer provided inside `\boxed{}` into `<answer>` tags, discarding the justification part of the answer, while keeping the thinking part of the dataset the same.

The results of both adjustments can be found in Table 4. We observe that training on only natural images in VLAA-Thinking hurts performance further, likely due to the limited data diversity. On the other hand, when applying improved answer formatting the results on Virgo improve slightly from 50.87% to 52.58%. However, despite these adjustments, the datasets still fail to improve beyond the base model.

E Full MMLU-Pro Evaluation Results

We provide the detailed results on all MMLU-Pro categories in Table 5. We observe that the model fine-tuned on our LongPerceptualThoughts dataset consistently outperforms the baselines and provides improvements on top of the base model except for the Other category.

F Implementation Details

F.1 LongPerceptualThoughts

Data generation. Our framework consists of three stages: generates verifiable multiple-choice questions using \mathcal{M}_{LLM} , extracts simple chains of thought (CoTs) from vision-

language models \mathcal{M}_{VLM} , and expands them into rich, long-form reasoning traces using frontier reasoning models $\mathcal{M}_{\text{Reason}}$.

1. In Stage 1, we use gpt-4o-mini-2024-07-18 with temperature 0.7. First, we prompt GPT-4o using the prompt in Fig. 5 to generate multiple-choice questions. Then, we parse the outputs by prompting GPT-4o again using the prompt in Fig. 6.
2. In Stage 2, we use Qwen2.5-VL-7B-Instruct with temperature 0.7, top_p 0.8, repetition_penalty, 1.05, and set number of samples per input to 3
3. In Stage 3, we use R1-Distill-Qwen-32B with temperature 0.7, top_p 0.8, top_k 50, repetition_penalty, 1.05, and set number of samples per input to 3. To avoid outputs include phrases like "As the description says", we explicitly define bad_words as "describe, description, described, describes, descriptions, mention, mentions, mentioned, misread, text, stated, says, mental"

F.2 Training details

SFT Training. We fine-tune the language decoder with a batch size of 256, sweeping learning rates over $\{10^{-5}, 8 \times 10^{-6}, 6 \times 10^{-6}\}$. Training runs for up to 5 epochs with early stopping based on the average validation accuracy. We set the maximum image resolution to 512×512 and the input cutoff length to 1024.

DPO Training. We fine-tune the language decoder with a batch size of 256, sweeping learning rates over $\{1 \times 10^{-6}, 5 \times 10^{-7}, 1 \times 10^{-7}\}$. Training runs for up to 3 epochs with early stopping based on the average validation accuracy. We set the maximum image resolution to 512×512 and the input cutoff length to 1024. For DPO, we set β to 1. and following Pang et al. (2024), we include SFT loss with a weight of 0.5.

F.3 DOCCI Captions

We select the same 500 images used to generate our dataset. Next, we format the training dataset with the user prompt "Provide a detailed description of the image.", prepending the image token and use the dense description provided in the dataset as the target answer of the model without further processing. We train the model using learning rate 8×10^{-6} with batch size 256 for a maximum of 20 epochs. The training reaches maximum average accuracy on the validation set in the third epoch and we subsequently use this checkpoint to report results in the main paper.

F.4 VLAA-thinking

We preprocess the dataset into two different versions, discarding samples where no reasoning trace could be extracted. The first version uses 24,035 randomly selected samples from the original dataset containing 158,827 samples. The second version also 24,035 random samples, however, we filter the dataset for images from ALLaVA-LAION and VizWiz. The latter specifically contains natural images - similar to the setting we train and evaluate on. We use the official dataset³ provided and apply some minor processing to the dataset to format the samples into a similar format as ours. In particular, we extract the thinking process and the answer from the *ds.answer* column of the dataset and place these into <think> and <answer> tags respectively. We use the same system prompt as for our model (see Sec. F.7).

Training. We finetune the language decoder using batch size 256. For both versions, we perform hyper parameter tuning by sweeping learning rates $\{10^{-5}, 8 \times 10^{-6}, 6 \times 10^{-6}\}$. We train for a maximum of 5 epochs and perform early stopping based on the average accuracy on the validation datasets.

F.5 Virgo

We use the dataset introduced in Virgo (Du et al., 2025) as D_{SD} ⁴ as other versions are not publicly available and it provides the best average performance in their experiments. As

³<https://huggingface.co/datasets/UCSC-VLAA/VLAA-Thinking>

⁴<https://huggingface.co/datasets/RUC-AIBOX/Virgo-Visual-Long-Thought-Dataset>

instructed on the webpage we use the “conversation” column of the dataset which the authors report to be the final data used to train the Virgo-7B model. The conversation column is constructed as the correct response with the shortest length of 5 samples given each prompt.

We apply minor processing to the dataset to follow our format by replacing the `<|begin_of_solution|>` and `<|end_of_solution|>` with `<answer>` and `</answer>`. Similarly, we replace `<|begin_of_thought|>` and `<|end_of_thought|>` with `<think>` and `</think>`. Finally, we append “Format the answer with the letter of the correct option in parentheses.” to the system prompt if the question is a multiple choice question. Overall, the resulting training dataset contains 14,540 samples.

Training. For training, we follow the setup described in F.4, *i.e.*, performing basic hyperparameter tuning, with the only change to limit training to 3 epochs as we found that the model performance peaks early during training. Surprisingly, we achieve the best validation performance before the first epoch ends.

F.6 Evaluation

Inference setup. We use vLLM (Kwon et al., 2023) for inferencing all models with greedy decoding. Detailed settings can be found in Tbl. 6. Further, we resize images’ longer side to 512 pixels preserving the aspect ratio if necessary. As the reasoning traces for MMLU-Pro tend to be longer for all models due to the difficulty of the task, we double the maximum number of new tokens generated. We use four NVIDIA RTX6000.

| Setting | Value |
|----------------|--------------------------|
| cutoff_length | 2048 |
| max_new_tokens | 1024 (2048 for MMLU-Pro) |
| temperature | 0.0 |
| top_p | 1.0 |
| dtype | half |

Table 6: vLLM inference settings.

F.7 Training and Evaluation Prompts

We provide the prompts for training and evaluation:

1. Fig. 7: The prompt used to train VLMs on DOCCI descriptions.
2. Fig. 8: The prompt used to evaluate VLMs to provide direct answers.
3. Fig. 9: Inspired by the prompt provided by DeepSeek-R1 (DeepSeek-AI et al., 2025), we design the prompt used to evaluate VLMs to provide thoughts and answers.

G Qualitative dataset example

We provide an example of our dataset in Fig. 10.

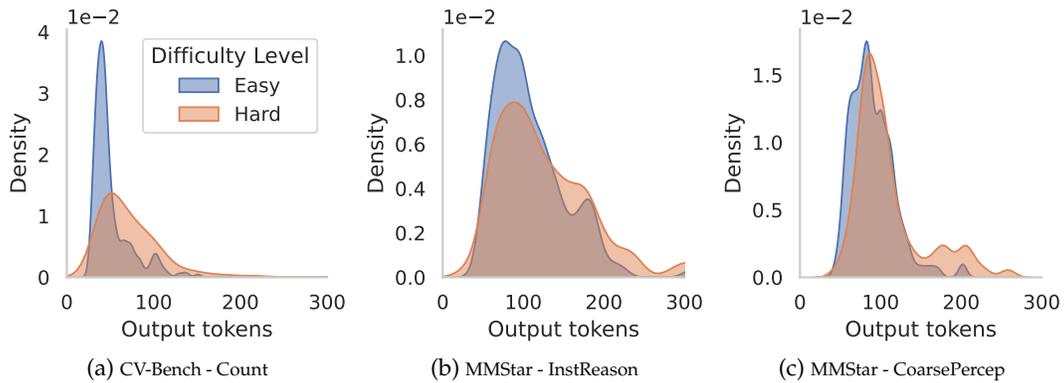


Figure 4: **Response lengths vs. question difficulties.** We analyze the responses of the VLM fine-tuned on LongPerceptualThoughts via DPO. Interestingly, we find that the model finetuned in our data naturally allocates more test-time compute for hard questions. We follow [Lightman et al. \(2024\)](#); [Snell et al. \(2025\)](#) and determine question complexity using rollouts on the base model.

System: You are an assistant that converts image descriptions to
↳ multi-choice visual questions.

User: Task:
You are given a detailed description of an image. Your goal is to
↳ generate diverse vision-centric, detailed questions that require a
↳ careful examination of the image for subtle visual details. Each
↳ question should be answerable in a brief sentence or single phrase.

Instructions:

- Focus on Visual Detail:
 - Ask questions that require examining fine details such as textures,
↳ patterns, and small or hidden elements.
 - Encourage the reader to analyze spatial relationships like object
↳ overlap, perspective, and layout.
 - Include aspects of lighting, shadows, and color gradients that
↳ affect the image's mood and depth.
- Comprehensive Coverage:
 - Ensure that the questions, as a group, address the majority of
↳ important details mentioned in the image description.
- Design for Multiple-Choice:
 - For each question, provide 4 answer options labeled A, B, C, and D.
 - Include one correct answer and three plausible distractors.
- Encourage Careful Inspection:
 - Design each question so that it cannot be answered without a close,
↳ careful visual inspection of the image.
 - Avoid generic or overly broad questions; each should target specific
↳ visual cues mentioned or implied in the description.
- Clarity, Specificity, and Brevity in Answers:
 - Formulate questions that are clear and focused on visual elements.
 - Ensure each question is detailed enough to challenge the reader to
↳ look beyond the surface.
 - Avoid questions that can be answered with general knowledge or
↳ assumptions.
 - Each question should be answerable in a brief sentence or even a
↳ single phrase.
- Structured Output:
 - Provide the questions in a numbered list.
 - Example Format: 1. <question> question here </question> <choices>
↳ (A) ... (B) ... (C) ... (D) ... </choices> <answer> short answer
↳ here </answer>

Image Description:
[IMAGE DESCRIPTIONS]
Assistant:

Figure 5: Text prompt converting descriptions to multi-choices questions.

```
User: You are given a text containing multiple multi-choice questions.
→ Each question includes a question statement, several choices, and an
→ answer. Your task is to reformat the text so that each multi-choice
→ question follows the structure below:

1. <question> question text here </question> <choices> (A) choice A text
→ (B) choice B text (C) choice C text (D) choice D text </choices>
→ <answer> answer text here </answer>

Please ensure that:
- Each question is numbered sequentially (e.g., 1., 2., 3., ...).
- The question portion is enclosed within the `<question>` tags.
- All answer options are enclosed within the `<choices>` tags, with each
→ option clearly labeled (A), (B), (C), (D).
- The answer is enclosed within the `<answer>` tags.
- The original content is preserved, but any formatting issues are
→ corrected according to the template above.

Here is the original content: [PREVIOUS_RESPONSE]
Assistant:
```

Figure 6: Text prompt to parse the response of Fig. 5 to multi-choices questions.

```
System: A conversation between User and Assistant. The user asks a visual
→ question, and the Assistant solves it. The answer are enclosed within
→ <answer> </answer> tags, respectively, i.e., <answer> answer here
→ </answer>. Format the answer with the letter of the correct option in
→ parentheses.
User: <image>Provide a detailed description of the image.
Assistant: [IMAGE DESCRIPTION]
```

Figure 7: Training prompt for training on DOCCI descriptions.

```
System: A conversation between User and Assistant. The user asks a visual
→ question, and the Assistant solves it. The answer are enclosed within
→ <answer> </answer> tags, respectively, i.e., <answer> answer here
→ </answer>. Format the answer with the letter of the correct option in
→ parentheses.
User: <image>[QUESTION]
Select from the following choices.
[CHOICES]
Assistant:
```

Figure 8: Evaluation prompt for direct answers.

System: A conversation between User and Assistant. The user asks a visual
→ question, and the Assistant solves it. The assistant first thinks
→ about the reasoning process in the mind and then provides the user
→ with the answer. The reasoning process and answer are enclosed within
→ `<think> </think>` and `<answer> </answer>` tags, respectively, i.e.,
→ `<think> reasoning process here </think> <answer> answer here`
→ `</answer>`. Format the answer with the letter of the correct option in
→ parentheses.
User: `<image>[QUESTION]`
Select from the following choices.
`[CHOICES]`
Assistant:

Figure 9: Evaluation prompt for thoughts and answers.



Question: How is the smaller tennis racket decorated within its strings? Select from the following choices.

- (A) A smiley face
 - (B) A floral design
 - (C) A logo
 - (D) A pattern of stars
- % Correct answer: (A) A smiley face

Simple CoT

<think>

Upon examining the smaller tennis racket, I can see that there is a
→ distinct pattern of stars within its strings. This pattern is
→ consistent across the strings, making it clear what the design is.
</think>

<answer> (D) </answer>

Ours

<think>

Upon examining the smaller tennis racket, I can see that there is a
→ distinct pattern of stars within its strings. This pattern is
→ consistent across the strings, making it clear what the design is.
→ Wait, no, hold on. Actually, the image shows a smiley face inside the
→ strings, but it's upside down. So that must be the decoration.

Wait, maybe I got confused. Let me double-check. The smaller racket has a
→ smiley face printed in black within the strings. So the correct
→ answer should be A smiley face. But wait, the options include both A
→ and C. Hmm, could it be a logo? No, because the Wilson racket has a
→ logo, but the smaller one has a smiley face. So the decoration is a
→ smiley face, which is option A.

</think>

<answer> (A) </answer>

Figure 10: Qualitative example of our LongPerceptualThoughts dataset performing verification.