VACoT: Rethinking Visual Data Augmentation with VLMs

Zhengzhuo Xu[1,2] Chong Sun[2] SiNan Du[1] Chen Li[2] Jing LYU[2] Chun Yuan[1]
[1]Tsinghua University [2]WeChat Vision, Tencent Inc.

Abstract

*While visual data augmentation remains a cornerstone for training robust vision models, it has received limited attention in visual language models (VLMs), which predominantly rely on large-scale real data acquisition or synthetic diversity. Consequently, they may struggle with basic perception tasks that conventional models handle reliably. Given the substantial cost of pre-training and fine-tuning VLMs, continue training on augmented data yields limited and diminishing returns. In this paper, we present **V**isual **A**ugmentation **C**hain-**o**f-**T**hought (**VACoT**), a framework that dynamically invokes image augmentations during model inference. By incorporating post-hoc transformations such as denoising, VACoT substantially improves robustness on challenging and out-of-distribution inputs, especially in OCR-related adversarial scenarios. Distinct from prior approaches limited to local cropping, VACoT integrates a structured collection of general visual augmentations, broadening the query image views while reducing training complexity and computational overhead with efficient agentic reinforcement learning. We propose a conditional reward scheme that encourages necessary augmentation while penalizing verbose responses, ensuring concise and effective reasoning in perception tasks. We demonstrate the superiority of VACoT with extensive experiments on 13 perception benchmarks and further introduce AdvOCR to highlight the generalization benefits of post-hoc visual augmentations in adversarial scenarios.*

1. Introduction

With scaling laws empirically validated on Large Language Models (LLMs), Visual Language Models (VLMs) have also achieved breakthroughs in image perception [6, 58], reasoning [1, 51], and generation [7, 10] tasks. While VLMs outperform traditional deep models on tasks such as key information extraction [55] and security auditing [21], they are highly vulnerable to adversarial samples. Fig. 1 shows that slight perturbations can mislead VLM perception results, enabling harmful content to evade system detection. Hence, ensuring adversarial robustness of VLMs remains an open and challenging problem.

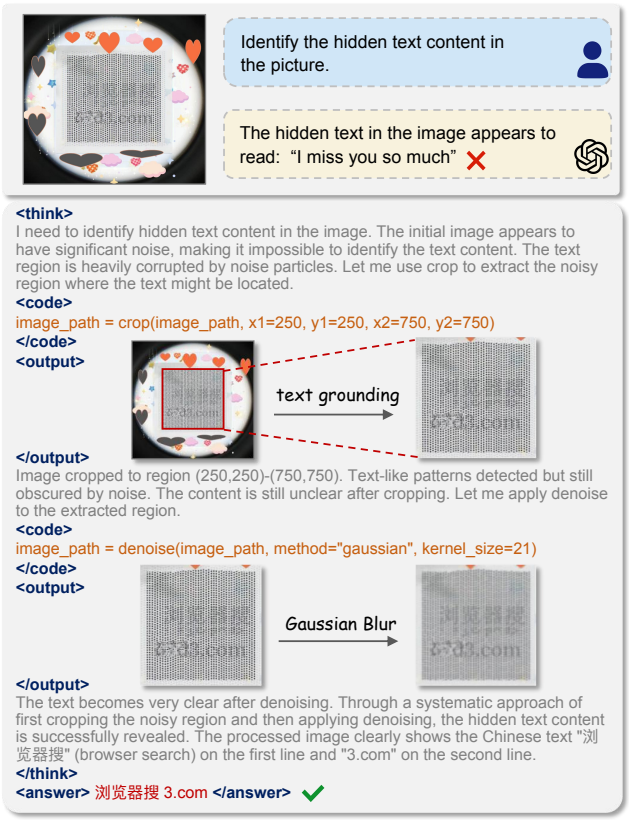

Figure 1. VACoT inference example. We address adversarial text recognition through iterative post-hoc visual augmentations.

Previous models rely on extensive visual data augmentations [12, 26, 60] to improve perception robustness. However, such augmentation introduces limited additional multimodal knowledge while substantially increasing computational costs during Pre-Training (PT) and Supervised Fine-Tuning (SFT). Hence, existing VLMs prioritizes large-scale real-world data or diverse domain-specific synthetic data, instead of more augmented visual questions. Recent studies [45, 48, 62, 67, 69] have explored the *thinking with bounding boxes* paradigm, which employs local cropping to focus model attention on specific regions. However, this strategy constitutes merely a specific instance of visual information filtering. More broadly, we advocate for *thinking with augmentation*, reformulating cropping and other

visual augmentations as post-hoc processing steps. This enhances model perceptual robustness while avoiding the training overhead caused by redundant image augmentations. As illustrated in Fig. 1, we propose Visual Augmentation CoT (VACoT) based on chat history concatenation. It stops autoregressive generation at designated tokens to either produce visual augmentations or terminate responses. By re-integrating returned information or augmented images into the conversation, we adopt end-to-end optimization via agentic Reinforcement Learning (RL). We wrap all augmentations into lightweight API calls to reduce the training difficulty and avoid uncontrollable actions.

Existing instruction-based agents struggle to comprehend when and what to apply visual augmentations based solely on textual descriptions. Hence, VACoT adopts a three-stage training pipeline to overcome it. *Stage* 1: We employ knowledge SFT with a difficulty-based data filtering strategy to efficiently enhance the model's foundational capabilities. *Stage* 2: We employ format SFT for visual augmentation cold-start initialization. Generating reliable trajectory data for post-hoc augmentation requires substantial manual effort and cost. Hence, we prompt the teacher models to deliberately insert random API calls when rewriting answers. This stage focuses on learning the correct calling format, even if these calls are not semantically relevant to the current query. *Stage* 3: We perform end-to-end agentic reinforcement learning with carefully designed reward signals, enabling the model to adaptively determine when and which augmentations to apply. We employ Qwen3 [59] as a teacher model to provide reward signals for verifying both answer correctness and formatting consistency. We penalize unnecessarily long reasoning traces that contribute limited performance gains. To this end, we incorporate the consistency reward [66] to assess the reasoning process, and further introduce a novel conditional API-call reward that promotes effective visual augmentation while preventing sequence explosion caused by indiscriminate trials.

Extensive evaluations on 13 public benchmarks demonstrate that our post-hoc visual augmentation significantly enhances perceptual capabilities, particularly on tasks requiring fine-grained recognition or robustness against adversarial text. We further introduce AdvOCR, a challenging benchmark comprising 100 adversarial samples for perception evaluation. Although it focuses solely on perception tasks, the benchmark is particularly challenging because of the adversarially modified images. Our VACoT could respond directly to clear queries and leverage iterative augmentations on adversarial samples to achieve high-confidence answers. In summary, our key contributions are:

a) We propose VACoT that integrates post-hoc visual augmentations into agentic workflows, significantly enhancing VLM robustness against adversarial inputs.
b) We devise an efficient three-stage training pipeline to learn adaptive visual augmentation strategies without prohibitive manual annotation.
c) We design a conditional API-call reward to effectively promote necessary augmentations while penalizing redundant actions in reinforcement learning.
d) We present a challenging benchmark called AdvOCR to rigorously evaluate the perceptual robustness of VLMs with fine-grained or adversarial images.

2. Related Works

Vision Language Models. Research has advanced from sophisticated projectors like Q-Formers [2, 18] for modality alignment to more streamlined designs utilizing linear projectors coupled with instruction tuning [23, 57]. For model ecosystems, proprietary models like the GPT [1] and Claude [38] series represent the state-of-the-art in performance. Meanwhile, the open-source community has been significantly advanced by series such as LLaMA [43, 44] and its derivatives (e.g., LLaVA [23, 24]), with recent series like QwenVL [3, 4, 46] and InternVL [5, 8, 63, 70] pushing the performance frontier. The DeepSeek [22, 54] and Mistral [13] series have further contributed by exploring scalable architectures like mixture of experts.

Visual Reasoning. CoT [50] and its variants [32, 68] enhance reasoning by decomposing complex problems into sequential steps. This paradigm has been significantly advanced by models like GPT-o1 [1] and DeepSeek-R1 [11], which employ reinforcement learning to bolster reasoning capabilities. Performance has also been improved by enabling longer output sequences directly [31, 33, 37, 51]. Subsequent efforts have explored multimodal CoT, where the reasoning structure is dynamically refined across multiple generation turns by incorporating region cropping [68] or external knowledge [17] to bolster interaction with visual tokens. Recent work [45, 52, 62, 67, 69] focuses on end-to-end training for tool invocation, aiming to equip models with autonomous multimodal reasoning. Nevertheless, existing methods are still confined to local region positioning for addressing fine-grained recognition.

Perception Benchmarks. Existing perception benchmarks primarily focus on chart understanding [27, 28, 49, 56] and document recognition tasks [14, 25, 29, 30, 36]. However, state-of-the-art models have largely saturated these benchmarks, achieving performance exceeding 90%. While recent benchmarks [9, 34] have increasingly emphasized reasoning complexity, the perceptual aspect of textual content in images remains relatively basic. In contrast, AdvOCR is derived from real-world failures and is designed to assess robust text perception in adversarial scenarios.

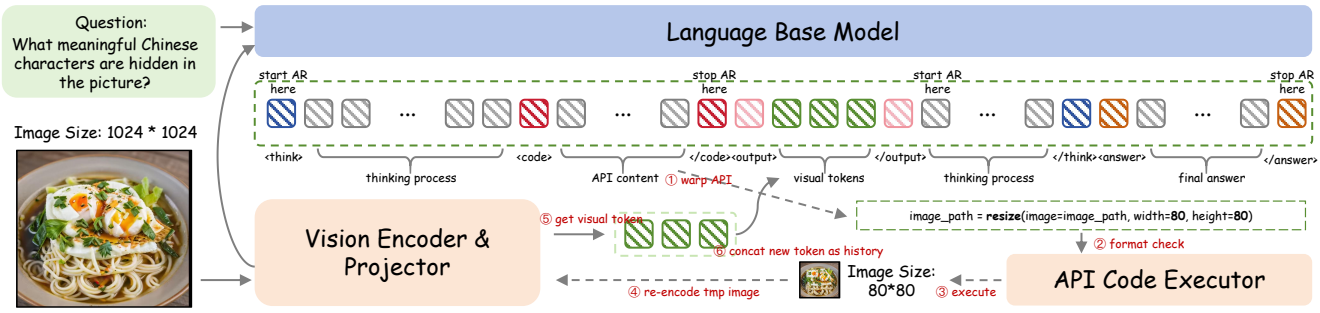

Figure 2. The overall architecture of VACoT. We leverage stop-words to achieve iterative post-hoc visual augmentation, providing more diverse image perspectives and higher-quality visual interactions compared to cropping-only agentic models.

3. Method

3.1. VACoT

Notation. Let I^0 be the query image and encoder $E(\cdot)$ outputs visual tokens $v^k = E(I^k)$ for the k-th state. Base LLM $L(\cdot)$ generates token sequence $y_{1:t}$ up to t. Available visual augmentation set is $\mathcal{A}$ and state k action is $a^k \in \mathcal{A} \cup \{\varnothing\}$. $\mathrm{Exec}(I, a)$ applies a to I to return the transformed image or error message e. H_t is the full context at step t, and κ is the generate stop token set.

Visual Augmentation CoT. Fig. 2 shows the architecture of VACoT, which is based on Qwen2.5VL-3B [4]. At reasoning step t^k, the language model $L(\cdot)$ autoregressively generates tokens conditioned on the current context H_{t-1}:

$$P(y_t \mid H_{t-1}) = L(y_t \mid H_{t-1}). \tag{1}$$

When the model emits token y_t^k in κ, the token sequence is parsed into a candidate operation $\hat{a}^k$ via a syntax parser $\hat{a}^k = \texttt{parse}(y_{t^{k-1}} : y_{t^k})$, where $\hat{a}^k$ represents the span of tokens enclosed by the `<code></code>` tag. The selected operation is executed through the API code executor:

$$e^k, I^k = \mathrm{Exec}(I^{k-1}, a^k). \tag{2}$$

The augmented image I^k is re-encoded by the vision encoder $E(\cdot)$ to obtain new visual tokens:

$$h^k = \begin{cases} v^k = E(I^k), & \text{if } I^k \text{ is available} \\ e^k, & \text{otherwise} \end{cases} \tag{3}$$

The chat history is updated with new visual embedding v^k or execution error message e^k as follows:

$$H_t = H_{t-1} \oplus \texttt{<output>}h^k\texttt{</output>}, \tag{4}$$

where $\oplus$ denotes message concatenation. This closed-loop process enables iterative token generation, visual augmentation, and feedback integration into subsequent steps.

Augmentation Design. We first implement several augmentations as API calls (e.g., `brightness`, `contrast`, `saturation`, `sharpening`, and `thresholding`). However, manual evaluations reveal that these post-hoc operations have a negligible impact on the model's perceptual capability. Consequently, we refine the set to more stable operations for $\mathcal{A}$: `crop`, `resize` (↑/↓), `rotate`, `flip`, `denoise` (filtering), and `edge` (edge detection). API calls are only executed if the operation and its parameters match. Otherwise, the corresponding error message is returned. The call success rate is ensured via supervised fine-tuning, reward design, and rule-based re-checks, which prevent token waste and uncontrolled code generation.

3.2. Agentic RL

Strategy. We employ the on-policy GRPO algorithm [35] to guide baseline to effectively utilize visual augmentation tools. The core improvement lies in handling the trajectory token discrepancies introduced by multiple rounds of tool invocations. The i-th complete generation trajectory of VACoT is formulated as:

$$s_i = \{(v^0, s_i^Q) \oplus \sum_k^{\oplus} (s_i^k, a_i^k, h_i^k) \oplus s_i^A\}, \tag{5}$$

where s denotes the token sequence, with superscripts Q and A representing the initial query and the final answer, respectively. The operator $\sum^{\oplus}$ indicates the sequential concatenation of grouped elements. Accordingly, the token sequence used for loss computation is defined as:

$$\tau_i = \{s_i^Q \oplus \sum_k^{\oplus} (s_i^k, a_i^k) \oplus s_i^A\}. \tag{6}$$

Given a generation policy π_θ parameterized by θ and a rollout set G, the optimization objective is:

$$\mathcal{J}(\theta) = \frac{1}{\sum_{i=1}^{G} |\tau_i|} \sum_{i=1}^{G} \sum_{j=1}^{|\tau_i|} \sigma(r_{i,j}) - \beta \cdot \mathbb{D}_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}), \tag{7}$$

where r is trajectory reward, $\sigma(\cdot)$ is intra-trajectory normalization, and β controls the KL-divergence regularization.

Reward. For each trajectory s_i, we define 5 rewards with reward model Qwen3-30B-A3B [59]: **1)** R_{vqa}: We evaluate

the correctness of the reasoning and final answer based on the last 500 characters of s_i. R_{vqa} is continuous in [0, 1]. **2)** R_{fmt}: We use regex-based matching to detect `<think>` and `<answer>` tags [11]. R_{fmt} is binary in 0 or 1. **3)** R_{cst}: We evaluates whether the reasoning process contains redundant repetition and whether the final answer is logically consistent with the reasoning. R_{cst} is continuous in [0, 1]. **4)** R_{api}: We detect all API calls in s_i via regex and assess the validity of the operation names and parameters. R_{api} is binary in 0 or 1. **5)** R_{suc}: This reward is re-weighted by R_{vqa}, API calls times k in s_i and maximum allowed times K, which is:

$$R_{\text{suc}} = \begin{cases} 0, & \text{if } R_{\text{vqa}} < 0.5, \\ 1, & \text{if } R_{\text{vqa}} \geq 0.5, k \leq 2, \\ 1 - \dfrac{k-2}{K-2}, & \text{if } R_{\text{vqa}} \geq 0.5, k \in (2, K], \\ 0, & \text{if } R_{\text{vqa}} \geq 0.5, k > K. \end{cases} \tag{8}$$

This reward discourage excessively long reasoning or frequent tool calls in perception-oriented tasks. The final reward is weighted ($\hat{R}$) as follows:

$$r_i = \hat{R}_{\text{vqa}} + \hat{R}_{\text{fmt}} + \hat{R}_{\text{cst}} + \hat{R}_{\text{api}} + \hat{R}_{\text{suc}}, \tag{9}$$

where the weights we adopt are [1, 0.25, 0.5, 0.25, 0.5].

3.3. Training.

Stage 1: Knowledge Enhancement. Given that the baseline model has undergone extensive pre-training, we apply a difficulty-based data filtering strategy to improve knowledge SFT efficiency. First, we collect approximately 4M high-quality open-source QA pairs [14, 19, 27, 29, 61]. As illustrated in Fig. 3, we perform *pass@4* inference with the Qwen2.5-VL-3B model using vLLM [15] and high generative diversity. Qwen3-30B-A3B [59] evaluates each answer to assign the difficulty score according to the number of error answer. We randomly select 10% of the difficulty level 0 samples while retaining all samples with difficulty levels between 1 and 3. For samples rated at difficulty level 4, we employ Qwen2.5-VL-72B [4] to verify the semantic validity of each QA pair and discard those deemed inherently unanswerable. After filtering, we obtain a total of 411K samples for knowledge-enhanced SFT.

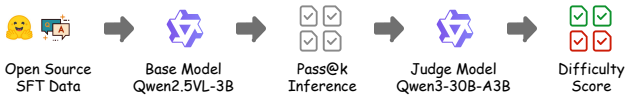

Figure 3. Illustration of pass@k data filter. The difficulty score is aligned with the number of correct answers for pass@k inference.

Stage 2: API Format. We rewrite model responses with Qwen2.5-VL-72B by inserting API call instructions to open-sourced data. As Fig. 4 shows, the `flip` and `rotate` adopt OCR data [27, 29, 65] with pre-flipped/rotate images; the `crop` and `resize` (↑) are applied to fine-grained recognition datasets [53, 67, 69]; and the `resize` (↓), `denoise`, and `edge` are applied to AIGuard [64] and synthetic hidden-text images [16]. Since collecting data that perfectly aligns with specific augmentations is challenging, we do not enforce strict consistency between augmentations and answers. This stage primarily focuses on format rather than whether the APIs are being invoked effectively.

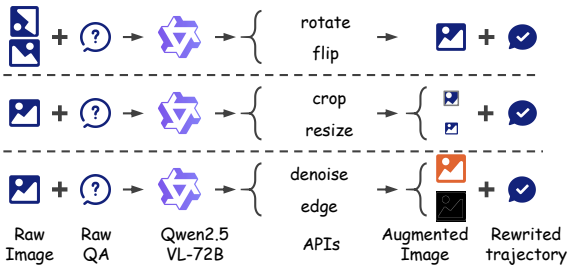

Figure 4. Construction of visual augmentation trajectory data.

Stage 3: Agentic RL. This stage trains the model not only to produce well-structured outputs but also to decide when and which API to invoke. This constitutes the key distinction between our approach and conventional instruction-following agents. We create a high-quality dataset of 66.4K samples, all suitable for post-hoc visual augmentation with high difficulty. The dataset spans four categories: fine-grained [67, 69], AIGC-generated [64], adversarial (from real-world), and OCR-hard samples (from stage 1).

4. AdvOCR

Motivation. Despite achieving near-saturated performance on standard perception benchmarks, frontier VLMs remain highly susceptible to carefully crafted adversarial images in real-world scenarios (Fig. 5). Such adversarial samples can easily mislead VLM-based systems into producing inappropriate or unsafe content. Hence, we introduce AdvOCR, a benchmark designed to evaluate the robustness of VLMs in perceiving adversarial textual content within images.

Construction. We collect the initial 1,241 images from manually reviewed bad cases and classic examples from the internet. We sequentially remove images that can be correctly recognized by both Qwen2.5-VL-72B [4] and GPT-5 [1]. The remaining ∼200 images are then manually categorized, from which we select the most representative samples in each category. To ensure data safety and diversity, we anonymize all sensitive information and synthesize additional adversarial samples based on the observed patterns to construct the final dataset.

Annotation. We design queries for all images following four principles: 1) each question should involve the perception of non-trivial visual elements; 2) the answer must be factually grounded in the image, without requiring complex reasoning; 3) each question should have a single, un-

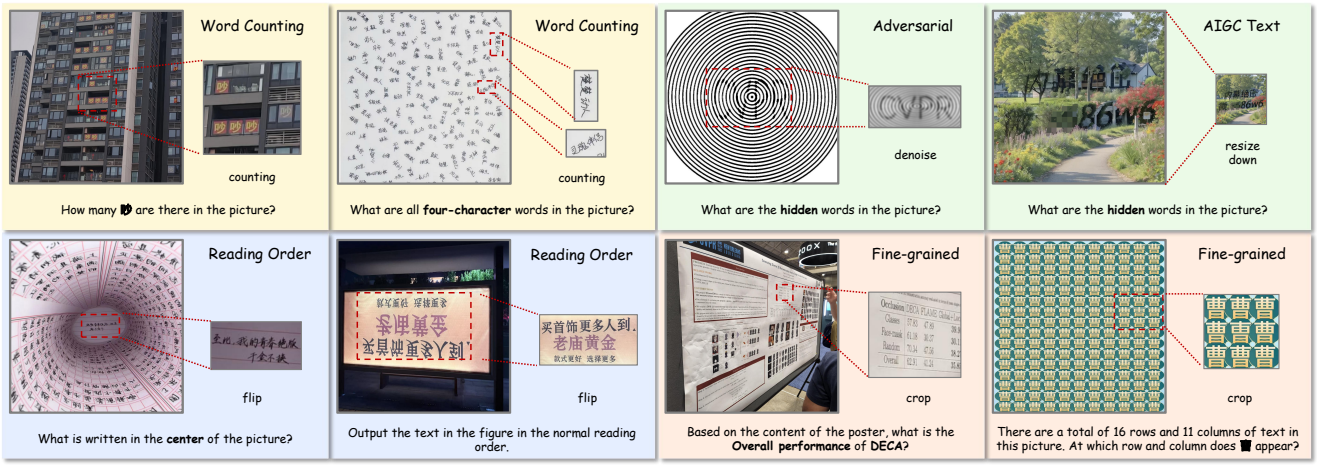

Figure 5. Example visualizations of AdvOCR. It poses greater demands on fine-grained and adversarial perception capabilities.

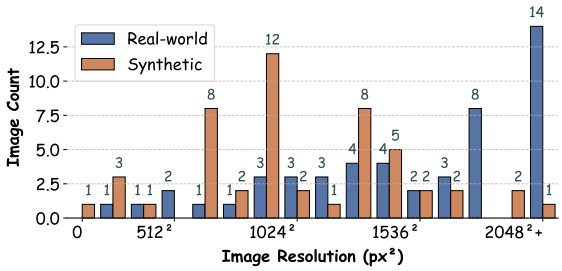

Figure 6. Image resolution distribution of AdvOCR.

ambiguous ground-truth answer; and 4) the question should be adversarially constructed to induce incorrect or uncertain responses from the GPT-5. This rigorous design process yielded 100 high-quality questions targeting challenging aspects of VLM perception.

Figure 7. Question type distribution of AdvOCR.

Statistic. AdvOCR consists of 100 manually designed adversarial OCR questions. Refer to Fig. 5 for representative examples. As shown in Fig. 6, the dataset exhibits a broad resolution distribution. The real-world split contains more high-resolution images for fine-grained visual recognition. Fig. 7 illustrates the question type distribution across 7 OCR-oriented tasks for both English and Chinese.

5. Experiment

5.1. Settings

Training Details. Our VACoT-3B is based on Qwen2.5VL-3B [4] with a 3-stage training pipeline. During the agentic RL stage, we employ GRPO for 2 epochs across 64 NVIDIA H20 GPUs. We set the maximum context length as 10,240 tokens and the completion length limit as 3,196 tokens to prevent excessive API calls. We use 4 rollout candidates for policy exploration, with sampling parameters of temperature = 1.0, $top_p = 0.9$, and $top_k = 50$. The total compute cost amounts to approximately 4,600 H20 GPU hours. Refer to Appendix A for more training details.

Baselines. Given that our model is built upon Qwen2.5VL-3B [4], the latter serves as the natural architectural baseline for our comparisons. We select InternVL3-2B [70] and OCRFlux-3B [39] for scale-comparable comparison. We further evaluate larger models such as Qwen2.5-VL-7B [4] and Thyme-7B [67]. Thyme employs executable code generation for tool reasoning, providing an alternative approach to our methodology. We reproduce Thyme-3B by the open-source codebase and training data for direct comparison at similar parameter scales. For proposed AdvOCR, we conduct extensive evaluation across local deployment models (Qwen2.5-VL-3B/7B/72B-Instruct [4] and Qwen3-VL-4B/8B-Instruct [59]) and proprietary API models, including GPT-4/4.1/4.1-mini/5/5-mini/o3 [1], Claude-Sonnet-4[38], Gemini-2.5-pro[6], Hunyuan-Turbos/Large-vision [42], Qwen3-VL-plus-250923 [41] and Doubao-Seed-1.6-vision-250815 [40].

Benchmarks. We conduct evaluations across three categories of perception benchmarks: 1) *General OCR*. This category encompasses chart-oriented benchmarks including AI2D [14], ChartQA [27], ChartQAPro [28], and $\text{CharXiv}_{\text{DQ}}$ [49], alongside document-oriented benchmarks comprising TextVQA [36], DocVQA [29], OCR-Bench [25], and InfoVQA [30]. 2) *Fine-grained Perception*. We employ HC-Bench [20] for hidden text detection, HRBench [47] for high-resolution content understanding, MME-Realworld [65] for real-world perception robustness, and Vstar [53] for spatial relationship comprehension. 3) *Proposed AdvOCR*. This benchmark integrates diverse challenging scenarios including text counting, hidden text perception, fine-grained visual recognition, and adversarial OCR samples. For evaluation consistency, we employ

Table 1. Top-1 accuracy (%) performance on 8 general perception-oriented benchmarks.

Benchmark	VACoT 3B	Qwen2.5VL [4] 3B	Thyme [67] 3B	InternVL3 [70] 2B	OCRFlux [39] 3B	Qwen2.5VL [4] 7B	Thyme [67] 7B
Chart-oriented							
AI2D [14]	**84.5**	78.2	80.8	78.7	80.6	83.9	84.2
ChartQA [27]	85.7	84.0	80.4	80.2	83.2	**87.3**	86.1
ChartQAPro [28]	**53.8**	41.7	43.1	-	41.0	51.8	52.3
CharXiv$_{DQ}$ [49]	**63.5**	56.2	57.4	54.7	55.0	60.4	62.3
Document-oriented							
TextVQA [36]	**85.3**	79.1	76.8	77.0	79.6	84.9	84.4
DocVQA [29]	95.6	93.9	92.2	88.3	82.4	**95.7**	95.3
OCRBench [25]	**87.7**	82.4	84.2	83.5	82.4	86.4	86.3
InfoVQA [30]	82.1	77.1	79.8	66.1	80.4	**82.6**	82.3
Average	**78.1**	72.2	71.8	75.8	70.3	77.3	77.4

Table 2. Top-1 accuracy (%) performance on 5 fine-grained or adversarial benchmarks.

Benchmark	Split	VACoT 3B	Qwen2.5-VL [4] 3B	Thyme [67] 3B	InternVL3 [70] 2B	OCRFlux [39] 3B	Qwen2.5-VL [4] 7B	Thyme [67] 7B
HC-Bench [20]	Object	**48.2**	0.9	1.8	0.0	0.9	0.9	3.6
	Text	**74.1**	4.5	1.8	5.4	3.6	3.6	0.9
	Overall	**61.2**	2.7	1.8	2.7	2.2	2.2	2.2
HR-Bench [47]	4K	74.8	63.0	70.1	57.0	61.5	67.6	**77.0**
	8K	70.6	59.5	65.4	48.9	60.9	60.4	**72.0**
	Overall	72.7	61.3	67.8	52.9	61.2	64.0	**74.5**
MME-Realworld$_{EN}$ [65]	Perception	**68.8**	57.1	62.1	49.1	54.1	63.0	67.1
	Reasoning	45.8	38.8	43.1	32.5	34.5	42.5	**48.4**
	Overall	**66.0**	54.9	59.8	47.1	51.7	60.6	64.8
MME-Realworld$_{CN}$ [65]	Perception	**71.9**	60.2	62.4	44.9	45.7	64.4	70.5
	Reasoning	**52.7**	42.0	46.4	36.6	29.7	42.5	52.1
	Overall	**65.7**	54.7	57.2	42.2	40.6	57.3	64.6
Vstar [53]	Attribute	**84.4**	80.9	80.9	72.2	80.0	79.1	83.5
	Spatial	**80.3**	63.2	75.0	71.1	63.2	75.0	80.3
	Overall	**82.7**	73.8	78.5	71.7	73.3	77.5	82.2

Qwen3-30B-A3B-2507[59] as judge model for all benchmarks, except for DocVQA and InfoVQA, which utilize official online evaluation platforms to ensure comparisons.

5.2. Main Results

Comparison on OCR Benchmarks. As shown in Tab. 1, we present comparisons with mainstream multimodal models (Qwen2.5VL [4], Thyme [67], InternVL3 [70], and OCRFlux [39]) on perception benchmarks. VACoT achieves a significant performance gain with an average score of 78.1% across eight benchmarks, marking a +5.9% improvement over the baseline model (Qwen2.5VL-3B at 72.2%). Notably, VACoT significantly outperforms not only models of similar scale but also larger ones like Qwen2.5VL-7B (77.3%) and even competes with Thyme-7B (77.4%). The improvements are especially pronounced on more challenging tasks: +12.1% on ChartQAPro (53.8% vs. Qwen2.5VL-3B 41.7%), +7.3% on CharXiv (63.5% vs. 56.2%), and +5.3% on OCRBench (87.7% vs. 82.4%). Combined with the tool invocation frequency shown in Tab. 6, Tab. 1 demonstrates the effectiveness of the knowledge SFT stage.

Comparison on Fine-grained Benchmarks. Tab. 2 shows comparisons on fine-grained visual benchmarks, which demand exceptional visual perception. On HR-Bench, VACoT-3B achieves 72.7%, significantly outperforming similar scale models and even matching the larger Thyme-7B. It attains 66.0% on MME-Realworld$_{EN}$ and 65.7% on MME-Realworld$_{CN}$, demonstrating robust, cross-lingual real-world understanding. VACoT achieves 82.7% on Vstar, a benchmark for attribute and spatial reasoning, narrowly yet significantly exceeding Thyme-7B 82.2%. HC-Bench requires models to identify hidden information in AIGC images. VACoT achieves 48.2% on object and 74.1% on text splits, dramatically outperforming all comparable models. In summary, VACoT achieves an average performance gain of over 10%, demonstrating the effectiveness of the proposed post-hoc visual augmentation.

Comparison on AdvOCR. AdvOCR comprises *real-world* and *synthetic* splits, designed to include challenging real-scenario OCR samples and adversarially crafted examples from observed failure patterns, respectively. Due to the high difficulty, Tab. 3 employs the pass@5 metric (correct if any of 5 attempts succeed). We use separate hyperparameter configurations: pass@1 (temp=0.1, top_p=0.8) and pass@5 (temp=0.7, top_p=0.95). Most models show limited improvement from pass@1 to pass@5, revealing a performance ceiling and constrained exploration space. Refer to Appendix B for case study. Models generally perform slightly better on the *real-world* split than the *syn-*

Table 3. Pass@k accuracy performance on proposed AdvOCR. All results are sorted according to pass@5 in the average split.

Model	Real-world			Synthetic			Average		
	pass@1	pass@5	Δ	pass@1	pass@5	Δ	pass@1	pass@5	Δ
Proprietary Models									
Claude-Sonnet-4 [38]	6%	10%	+4%	0%	0%	+0%	3%	5%	+2%
GPT-4-turbo [1]	0%	6%	+6%	10%	10%	+0%	5%	8%	+3%
GPT-5-mini [1]	10%	12%	+2%	16%	20%	+4%	13%	16%	+3%
GPT-4.1-mini [1]	10%	18%	+8%	10%	20%	+10%	10%	19%	+9%
Hunyuan-Turbos-vision [42]	16%	20%	+4%	18%	18%	+0%	17%	19%	+2%
GPT-5 [1]	12%	16%	+4%	22%	26%	+4%	17%	21%	+4%
Hunyuan-Large-vision [42]	18%	20%	+2%	**26%**	26%	+0%	22%	23%	+1%
GPT-4.1 [1]	18%	20%	+2%	**26%**	**30%**	+4%	22%	25%	+3%
GPT-o3 [1]	14%	22%	+8%	16%	28%	+12%	15%	25%	+10%
Gemini-2.5-pro [6]	22%	38%	+16%	8%	14%	+6%	15%	26%	+11%
Qwen3-VL-plus-2025-09-23 [59]	**38%**	**46%**	+8%	10%	12%	+2%	**24%**	29%	+5%
Doubao-Seed-1.6-vision-250815 [40]	22%	42%	+20%	12%	18%	+6%	17%	**30%**	+13%
Open-source Models									
Thyme-3B [67]	18%	18%	+0%	4%	6%	+2%	11%	12%	+1%
Qwen2.5-VL-72B-Instruct [4]	20%	20%	+0%	10%	10%	+0%	15%	15%	+0%
Qwen2.5-VL-3B-Instruct [4]	20%	24%	+4%	6%	6%	+0%	13%	15%	+2%
Thyme-7B [67]	20%	22%	+2%	6%	8%	+2%	13%	15%	+2%
Qwen2.5-VL-7B-Instruct [4]	24%	30%	+6%	4%	4%	+0%	14%	17%	+3%
Qwen3-VL-2B-Instruct [59]	22%	30%	+8%	8%	10%	+2%	15%	20%	+5%
Qwen3-VL-4B-Instruct [59]	34%	40%	+6%	8%	10%	+2%	21%	25%	+4%
VACoT-3B	**62%**	**78%**	+16%	**48%**	**56%**	+8%	**55%**	**67%**	+12%

thetic split, because the latter's adversarial patterns are more challenging. SOTA proprietary models manage compound Chinese characters and fine-grained recognition more effectively in real-world scenarios. The Qwen and Doubao series show remarkable performance due to Chinese-specific optimisation. VACoT-3B achieves significant gains: +21% pass@1 over Qwen3-VL-plus (24%) and +37% pass@5 over Doubao-Seed-1.6-vision (30%), validating our proactive augmentation strategy. Our model sustains significant pass@5 gains by efficiently exploring multiple image augmentations using minimal output tokens, thereby achieving more robust image views.

5.3. Ablation Study

Train Strategies. Tab. 4 shows the ablation study of training stages (details in Sec. 3.3). Settings A and B demonstrate that difficulty-aware data cleaning significantly boosts baseline performance. Setting C shows that direct agentic RL training is quite challenging, as the augmentations are not frequently invoked. The comparison between settings D and G indicates that the format SFT is essential for cold-start initialization. RL training provides clear and consistent gains on fine-grained benchmarks and AdvOCR.

Reward Design. In Sec.3.2, we introduce the API format reward R_{api} and API success reward R_{suc} for the accurate and efficient identification of post-hoc augmentations. Fig. 8 (left) shows that both rewards evolve steadily during training. With format learning in Stage 2, R_{api} quickly converges to a steady value. In contrast, R_{suc} would only be awarded when the model both invokes valid API calls and produces the correct answer, leading to a more gradual increase throughout training. Therefore, it eventually converges to approximately 0.3. Fig. 8 (right) shows that vanilla training manner leads the model to exhaustively try different augmentations, resulting in unnecessarily long responses that hinder perception efficiency. With proposed rewards, the response length converges to a reasonable range after a brief initial increase, demonstrating the effectiveness of our reward strategy.

Table 4. Ablation study of different training strategies. *General* refers to the average value in Tab. 1.

Index	Stage1	Stage2	Stage3	General	HC-Bench	Vstar	AdvOCR
A	✗	✗	✗	72.2	2.7	73.8	13.0
B	✓	✗	✗	76.3	3.6	76.4	18.0
C	✗	✗	✓	74.1	18.8	74.1	15.0
D	✗	✓	✓	74.0	26.3	75.8	31.0
E	✓	✓	✗	76.6	8.9	75.4	21.0
F	✓	✗	✓	77.4	25.5	78.5	22.0
G	✓	✓	✓	**78.1**	**61.2**	**82.7**	**55.0**

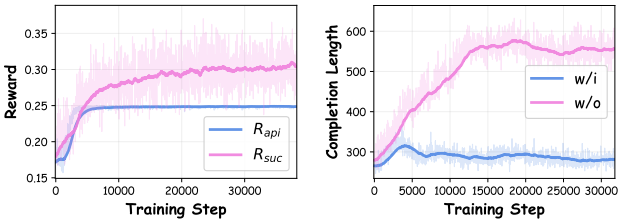

Figure 8. The EMA smooth curve of reward value and completion mean length during the agentic RL training.

5.4. Further Analysis

Comparison with Instruction-based Agent. We compare with manual post-hoc visual augmentation via pre-defined instructions. The model will be provided with tool descrip-

tions and forced to perform post-hoc augmentation before answering. We manually execute the first round of API calls and query the model with the augmented images. As shown in Tab. 5, this approach also consistently improves baseline performance across various models, highlighting the importance of post-hoc augmentation. However, the performance remains limited because they fail to select the most suitable augmentation efficiently. Conversely, when we explicitly prohibit VACoT from using API calls in the system prompt, performance drops significantly, confirming that the improvement stems from proper tool-based augmentation rather than other factors.

Table 5. Comparison with instruction-based agentic strategy on AdvOCR. Aug.: post-hoc visual augmentation.

Method	Aug.	Real-world	Δ	Synthetic	Δ	Average	Δ
GPT-5-mini	✗	10%	-	16%	-	13%	-
	✓	26%	+13%	20%	+4%	23%	+10%
Qwen2.5-VL-72B	✗	20%	-	10%	-	15%	-
	✓	32%	+12%	18%	+8%	25%	+10%
Ours	✗	28%	-34%	16%	-32%	22%	-33%
	✓	62%	-	48%	-	55%	-

Activate Frequency Statistics. Tab. 6 reports the frequency of API calls in VACoT's responses across multiple benchmarks. On OCRBench and ChartQA, the model can provide answers directly without invoking image transformations. In contrast, benchmarks containing adversarial or challenging samples exhibit a significantly higher rate of tool usage. Specifically, fine-grained recognition datasets trigger more frequent use of `crop` and `resize`, while adversarial samples lead to increased use of `resize` and `denoise`. The low failure rate demonstrates the effectiveness of our API-format SFT and reward design. For queries that exceed the maximum number of interaction attempts, the final response is generated by appending the following history: `OK, I have to give the final answer directly.`

Table 6. API call frequency (%) statistics. `direct`: no call is made. `fail`: invalid call or exceeds attempt limit.

Benchmark	OCRBench	ChartQA	HR-Bench	HC-Bench	AdvOCR
`direct`	67.3	42.4	16.7	13.8	13.0
`fail`	2.7	3.3	8.2	1.3	3.0
`crop`	15.3	44.5	72.2	12.3	14.0
`resize`	19.2	15.5	43.1	67.8	44.0
`flip`	3.3	1.8	6.8	12.5	12.0
`rotate`	8.8	9.2	0.4	2.4	7.0
`denoise`	9.2	4.4	4.2	14.7	38.0
`edge`	2.8	8.3	7.7	22.8	15.0

Visual Token Compression with Resize. Leveraging the post-augmentation, we explore input visual token compression via `resize`(↑). We conduct ablation study by downsampling input images to evaluate if lower resolutions suffice for certain queries. Images at the minimum resolution of 28×28 are excluded from this process. Our `resize`(↑) implementation is adapted such that if the target size is smaller than the original resolution, the `resize`(↑) will be based on the original image instead of interpolation on the downsampled one. As shown in Tab. 7, even with a 50% compression rate (equivalent to 75% reduction in visual tokens), performance drops by only ∼1% when `resize`(↑) is allowed. This resilience is attributed to the post-hoc resizing on ∼30% of uncertain cases, effectively recalling the enlarged image to preserve accuracy.

Table 7. Performance comparison with vs. without `resize` (↑) invocation on compressed-resolution images.

Benchmark	OCRBench			ChartQA		
Compress rate	100%	75%	50%	100%	75%	50%
`resize` (↑) rate	1.1%	14.2%	33.5%	0.0%	8.4%	33.2%
w/i `resize` (↑)	87.7	87.1	86.4	85.7	85.4	84.5
w/o `resize` (↑)	87.7	86.4	82.1	85.7	84.3	81.2

5.5. Discussion

Comparisons with Crop-based Methods. *Thinking-with-crop* paradigms localize regions before answering [45, 69]. However, cropping essentially represents a specific form of visual information filtering. In contrast, we generalize it to *thinking with augmentations*, where all augmentations are treated as diverse information filters executed during reasoning. Unlike code-based agents [67], we unify all augmentations into lightweight and deterministic API calls, making the outputs more controllable and enabling more iterative exploration under limited context length. Our conditional reward design discourages redundant transformations, improving both perceptual efficiency and robustness.

Inference Efficiency. Although VACoT introduces additional reasoning steps for post-hoc visual augmentations, the inference overhead remains acceptable. Fig. 8 shows that the conditional reward explicitly penalizes redundant or low-impact augmentations, leading the model to invoke augmentations only when necessary. The observations in Appendix B further reveal that the augmentation always begins by `crop` or `resize` (↓), which effectively limits the visual token sequence length introduced by subsequent multi-round augmentations.

6. Conclusion

We propose VACoT, a framework that enables dynamic invocation of image augmentation during inference for visual language models. Specifically, we unify a structured set of general visual augmentation strategies and introduce a conditional reward scheme to balance exploration and efficiency in RL training. We further contribute a challenging real-world OCR benchmark, AdvOCR, where VACoT achieves strong performance while state-of-the-art models fail without augmentation. Extensive experiments demonstrate the superiority of our approach.

Acknowledgement

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008), and Beijing Key Lab of Networked Multimedia.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 4, 5, 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A visual language model for few-shot learning. In *proceedings of NeurIPS*, pages 23716–23736, 2022. 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4, 5, 6, 7, 12

[5] Zheng Cai, Maosong Cao, et al. Internlm2 technical report. *arXiv preprint:2403.17297*, 2024. 2

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 5, 7

[7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1

[8] Xiaoyi Dong, Pan Zhang, Yuhang Zang, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint:2401.16420*, 2024. 2

[9] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024. 2

[10] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 1

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint:2501.12948*, 2025. 2, 4

[12] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mistral 7b. *arXiv preprint:2310.06825*, 2023. 2

[14] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 2, 4, 5, 6

[15] 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. 4

[16] Monster Labs. Controlnet qr code monster v2 for sd-1.5. https://huggingface.co/monster-labs/control_v1p_sd15_qrcode_monster, 2025. 4

[17] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, et al. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint:2501.07542*, 2025. 2

[18] Junnan Li, Dongxu Li, Silvio Savarese, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *proceedings of ICML*, pages 19730–19742, 2023. 2

[19] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 4

[20] Sifan Li, Yujun Cai, and Yiwei Wang. Semvink: Advancing vlms' semantic understanding of optical illusions via visual global thinking, 2025. 5, 6

[21] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025. 1

[22] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint:2405.04434*, 2024. 2

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *proceedings of NeurIPS*, 2023. 2

[24] Haotian Liu, Chunyuan Li, Yuheng Li, et al. Improved baselines with visual instruction tuning. In *proceedings of CVPR*, 2024. 2

[25] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2(5): 6, 2023. 2, 5, 6

[26] Yuliang Liu, Jiaxin Zhang, Dezhi Peng, Mingxin Huang, Xinyu Wang, Jingqun Tang, Can Huang, Dahua Lin, Chunhua Shen, Xiang Bai, et al. Spts v2: single-point scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15665–15679, 2023. 1

[27] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2, 4, 5, 6

[28] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025. 2, 5, 6

[29] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2, 4, 5, 6

[30] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2, 5, 6

[31] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, 2025. 2

[32] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *proceedings of CVPR*, pages 14420–14431, 2024. 2

[33] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025. 2

[34] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025. 2

[35] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[36] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2, 5, 6

[37] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025. 2

[38] Anthropic Team. Claude's extended thinking. https://www.anthropic.com/news/visible-extended-thinking, 2025. 2, 5, 7

[39] ChatDoc Team. Ocrflux-3b. https://huggingface.co/ChatDOC/OCRFlux-3B, 2025. Accessed: 2025-10-23. 5, 6

[40] Doubao Team. Seed1.6 tech introduction. https://seed.bytedance.com/en/seed1_6, 2025. 5, 7

[41] Qwen Team. Qwen chat. https://qwen.ai/qwenchat, 2025. 5

[42] Tencent Hunyuan Team. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. *arXiv preprint arXiv:2505.15431*, 2025. 5, 7

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint:2302.13971*, 2023. 2

[44] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint:2307.09288*, 2023. 2

[45] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025. 1, 2, 8

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[47] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7907–7915, 2025. 5, 6

[48] Ye Wang, Qianglong Chen, Zejun Li, Siyuan Wang, Shijie Guo, Zhirui Zhang, and Zhongyu Wei. Simple o3: Towards interleaved vision-language reasoning. *arXiv preprint arXiv:2508.12109*, 2025. 1

[49] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. 2, 5, 6

[50] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *proceedings of NeurIPS*, pages 24824–24837, 2022. 2

[51] Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025. 1, 2

[52] Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Ze-

jia Weng, et al. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *arXiv preprint arXiv:2507.05255*, 2025. 2

[53] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 4, 5, 6

[54] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint:2412.10302*, 2024. 2

[55] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357, 2024. 1

[56] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint:2312.15915*, 2023. 2

[57] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint:2408.08872*, 2024. 2

[58] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 3, 4, 5, 6, 7

[60] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022. 1

[61] Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, et al. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *arXiv preprint arXiv:2502.14846*, 2025. 4

[62] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025. 1, 2

[63] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint:2309.15112*, 2023. 2

[64] Wenhua Zhang, Weicheng Li, Xuanrong Rao, Lixin Zou, Xiangyang Luo, Chubin Zhuang, Yongjie Hong, Zhen Qin, Hengyu Chang, Chenliang Li, et al. Aiguard: A benchmark and lightweight detection for e-commerce aigc risks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12437–12450, 2025. 4

[65] 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. 4, 5, 6

[66] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025. 2

[67] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025. 1, 2, 4, 5, 6, 7, 8

[68] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, et al. Multimodal chain-of-thought reasoning in language models. *TMLR*, 2024, 2024. 2

[69] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1, 2, 4, 8

[70] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2, 5, 6