Zooming into Comics: Region-Aware RL Improves Fine-Grained Comic Understanding in Vision-Language Models

Master's thesis in Computer Systems & Networks

Yule Chen

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Zooming into Comics: Region-Aware RL Improves Fine-Grained Comic Understanding in Vision-Language Models

YULE CHEN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Computer and Network Systems
Chalmers University of Technology
Gothenburg, Sweden 2025

Zooming into Comics: Region-Aware RL Improves
Fine-Grained Comic Understanding in Vision-Language Models
YULE CHEN

Supervisor:
Sabine Süsstrunk & Yufan Ren, École Polytechnique Fédérale de Lausanne
Examiner:
Lars Hammarstrand, Department of Electrical Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Division of Computer and Network Systems
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Zooming into Comics: Region-Aware RL Improves Fine-Grained Comic Understanding in Vision-Language Models

Yule Chen
Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

Complex visual narratives, such as comics, present a significant challenge to Vision-Language Models (VLMs). Despite excelling on natural images, VLMs often struggle with stylized line art, onomatopoeia, and densely packed multi-panel layouts. To address this gap, we introduce **AI4VA-FG**, the first fine-grained and comprehensive benchmark for VLM-based comic understanding. It spans tasks from foundational recognition and detection to high-level character reasoning and narrative construction, supported by dense annotations for characters, poses, and depth. Beyond that, we evaluate state-of-the-art proprietary models, including GPT-4o and Gemini-2.5, and open-source models such as Qwen2.5-VL, revealing substantial performance deficits across core tasks of our benchmarks and underscoring that comic understanding remains unsolved. To enhance VLMs' capabilities in this domain, we systematically investigate post-training strategies, including supervised fine-tuning on solutions (SFT-S), supervised fine-tuning on reasoning trajectories (SFT-R), and reinforcement learning (RL). Beyond that, inspired by the emerging "Thinking with Images" paradigm, we propose **Region-Aware Reinforcement Learning (RARL)** for VLMs, which trains models to dynamically attend to relevant regions through zoom-in operations. We observe that when applied to the Qwen2.5-VL model, RL and RARL yield significant gains in low-level entity recognition and high-level storyline ordering, paving the way for more accurate and efficient VLM applications in the comics domain.

Keywords: comics, machine learning, deep learning, large language models, multimodality, post-training, agentic reinforcement learning

Acknowledgements

Acknowledgements

I would like to express my sincere gratitude to all those who supported and guided me throughout the course of this master's thesis.
First and foremost, I would like to thank my supervisor, Prof. Sabine Süsstrunk, for her insightful feedback and continuous encouragement throughout the project. Her expertise and vision have greatly shaped the direction of my research.
I would also like to thank my examiner, Prof. Lars Hammarstrand, for his constructive comments and thoughtful evaluation of my work.
Most importantly, I am deeply grateful to my mentor, Yufan Ren, whose generous assistance, patient mentorship, and countless helpful discussions have been instrumental to the success of this thesis. His support went far beyond academic advice, and I truly appreciate the time and effort he dedicated to helping me grow as a researcher.
Finally, I would like to acknowledge the support of my peers, friends, and family, who have accompanied me on this journey with unwavering encouragement.

Yule Chen, Lausanne, August 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AI	Artificial Intelligence
GRPO	Group Relative Policy Optimization
IoU	Intersection over Union
MLLM	Multi-modal Large Language Models
LLM	Large Language Models
OCR	Optical Character Recognition
RL	Reinforcement Learning
RLVR	Reinforcement Learning with Verifiable Rewards
ViT	Vision Transformer
VLM	Vision Language Model
VQA	Visual Question Answering

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Variables

τ	response trajectory
m	the number of zoom-in tool invocations
R	total reward
R_{acc}	accuracy reward
R_{format}	format reward
R_{tool}	tool-usage reward
$R_{\text{tool-count}}$	tool-usage count reward
$R_{\text{tool-acc}}$	tool-usage accuracy reward

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivations

The integration of vision and language understanding is a central goal in artificial intelligence research. Vision-Language Models (VLMs), trained to process and align visual and textual information, have demonstrated remarkable success in tasks such as image captioning, visual question answering, and cross-modal retrieval. However, understanding complex visual narratives—such as comics—remains a less explored and more challenging problem.

The long-term objective of this research direction is to enable creative generation within the comic domain, particularly comic storytelling at the chapter level, thereby supporting applications such as accessible storytelling for visually-impaired readers. Nevertheless, our preliminary experiments with state-of-the-art VLMs for page-level caption generation (see Appendix A.3) reveal several fundamental limitations: (1) Generated descriptions frequently lack fine-grained details and occasionally introduce factual inconsistencies (e.g. conflating distinct characters across panels) (2) All tested models exhibit difficulty in capturing inter-panel dependencies, resulting in narratives that fail to preserve the temporal coherence characteristic of comics. (3) Robust evaluation methodologies for comic generation remain absent; existing practices [4], which rely on stronger VLMs as automatic evaluators, are problematic in this setting, as these models themselves exhibit poor comic comprehension, leading to biased or unreliable quality assessments.

Motivated by these challenges, we direct our efforts toward the foundational problem of *comic understanding*. Accordingly, this thesis introduces a benchmark specifically designed to evaluate and advance VLMs' capacity to understand comics, especially to comprehend the distinctive narrative structures of comics.

1.2 Goals

The main goals of this thesis are summarized as follows:

- Develop **AI4VA-FG**, a fine-grained benchmark for comic understanding with vision-language models, covering both low-level recognition and high-level rea-

soning tasks with dense annotations for characters, poses, depth, and more.

- Benchmark state-of-the-art vision-language models on AI4VA-FG to evaluate their performance and identify common failure cases.
- Assess the effectiveness of post-training methods, including supervised finetuning (SFT) and reinforcement learning (RL), to improve model performance on comic understanding tasks.
- Design and implement **Region-Aware Reinforcement Learning (Region-Aware RL)**, a framework inspired by the "thinking-with-image" approach of OpenAI o3, which learns where and when to zoom for enhanced visual reasoning in comics.

1.3 Limitations / Demarcations

This research was carried out within a limited timeframe (March to July 2025), which constrained the breadth of experiments. Some design choices—such as model selection and parameter tuning—were made under tight deadlines, restricting the extent of iterative testing and optimization. Furthermore, with the rapid advancement of VLMs and the emergence of more efficient RL algorithms and higher-performing models on leaderboards, this study focuses on evaluating a representative subset of widely used models available during the research period.

1.4 Paper Structure

The remainder of this paper is organized as follows:

- **Chapter 2** provides background on MLLMs, including an overview of current progress in MLLM architecture design and post-training techniques, the domain of comics, and the use of Vision-Language Models (VLMs) for comic understanding.
- **Chapter 3** introduces **AI4VA-FG**, our newly proposed fine-grained benchmark for comic understanding, and presents evaluation results of state-of-the-art VLMs on this benchmark.
- **Chapter 4** details the methodology used to enhance VLM performance on AI4VA-FG, including SFT, RL, and our proposed Region-Aware Reinforcement Learning (Region-Aware RL) framework.
- **Chapter 5** presents the experimental results, along with in-depth analysis and discussions of model performance and common failure modes.
- **Chapter 6** concludes the thesis and outlines possible directions for future research.

2

Background

2.1 Multimodal Large Language Models (MLLMs)

The rise of large language models (LLMs) has transformed artificial intelligence, showcasing strong capabilities in understanding and generating human-like text through transformer-based architectures [5] trained on massive corpora. However, their text-only design limits their ability to interact with the inherently multimodal nature of the real world. Despite this, the success of LLMs in textual tasks provides a strong foundation for extending these architectures to handle diverse data modalities.

Vision-Language Models (VLMs) are a key subclass of Multimodal Large Language Models (MLLMs), designed to jointly process visual and textual inputs and generate primarily textual outputs. Their goal is to bridge the semantic gap between visions (images or videos) and language, enabling AI to understand the world through both modalities. The progress of VLMs has been driven by the availability of large-scale image-text datasets and the success of transformer-based architectures in vision and text tasks.

VLMs are meticulously designed to learn a shared, rich representation space where visual and textual information can be mutually understood and related, facilitating a wide array of cross-modal tasks. Representative early examples include CLIP[6], Flamingo [7], BLIP [8], and LLaVA [9], each with varying degrees and methodologies of visual and textual integration.

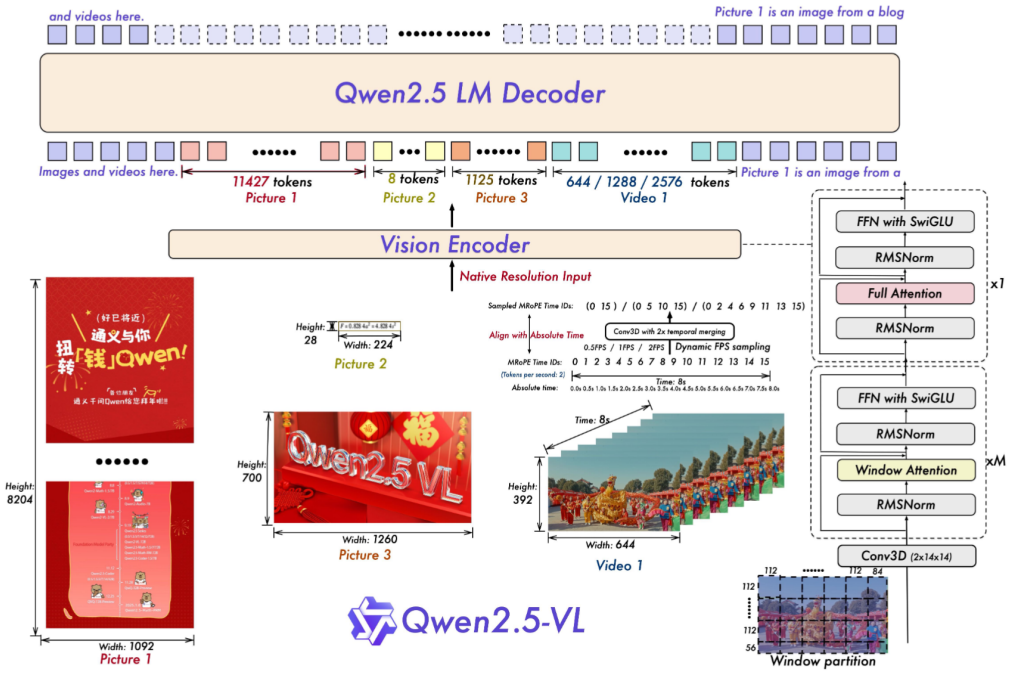

Figure 2.1: Example of state-of-the-art VLM architecture. The Qwen2.5-VL architecture presents the integration of a vision encoder and a language model decoder to process multimodal inputs [1].

2.1.1 Applications of VLMs

VLMs have transformed various domains by enabling systems to jointly process and reason over visual and textual information. Representative basic applications include:

- **Image Captioning:** Generating descriptive natural language for a given image. VLMs translate visual content into coherent, semantically meaningful text. Applications include assistance for visually impaired individuals, automated content indexing, and report generation.
- **Visual Question Answering (VQA):** Answering natural language questions based on image content. This requires both visual understanding and reasoning. For instance, given a comic panel and the question *"What is the main character doing?"*, the model must identify the character and determine its action.
- **Visual Grounding:** Identifying specific objects or regions (e.g. panels, characters) in an image based on a textual reference. This ability is crucial for tasks such as object detection, visual dialogue, and human-robot interaction.
- **Visual Reasoning:** Performing logical or relational inferences over visual

inputs, often guided by textual prompts. Examples include spatial reasoning tasks like determining whether *"the second panel in the bottom row has 2 characters"*, which require compositional understanding of the scene.

Building on these core capabilities, VLMs show strong potential for understanding complex multimodal content such as comics—an area that has not yet been extensively explored—enabling them to reason about and even generate captions for comic pages.

2.1.2 Model Selection

Our evaluation suite encompasses both proprietary and open-source MLLMs that represent the current state-of-the-art. For proprietary systems, given economic constraints, we select Gemini-2.5-pro [10] and GPT-4o [11] as leading commercial baselines. On the open-source side, given computational constraints, we primarily focus on models under 10B parameters, including Qwen2.5-VL-3B/7B/32B [1], and InternVL3-8B/9B [12]. Most other leading open-source models share the same LLM or ViT backbones as these representatives, making our selection broadly representative of current open-source progress. Notably, several small open-source models achieve comparable or even higher scores than GPT-4o on the Huggingface OpenVLM Leaderboard, which evaluates VLMs on widely used benchmarks such as MMBench [13] and MMMU [14].

Table 2.1: Summary of selected VLMs. Parameter sizes of the proprietary models are not publicly disclosed.

Model	Param (B)	Language Model	Vision Model	Score[1]
Qwen2.5-VL-3B	3.75	Qwen2.5-3B	QwenViT	64.5
Qwen2.5-VL-7B	8.29	Qwen2.5-7B	QwenViT	70.9
Qwen2.5-VL-32B	33.5	Qwen2.5-32B	QwenViT	74.8
InternVL3-8B	7.94	Qwen2.5-7B	InternViT-300M-v2.5	73.6
InternVL3-9B	9.14	InternLM3-8B	InternViT-300M-v2.5	72.6
MiniCPM-o-2.6	8.67	Qwen2.5-7B	SigLIP-400M	70.2
GPT-4o-2024-08-06	/	/	/	71.6
GPT-5-mini	/	/	/	/
Gemini-2.0-flash	/	/	/	72.6
Gemini-2.5-flash	/	/	/	/
Gemini-2.5-pro	/	/	/	80.9

[1] Huggingface OpenVLM LearderBoard.

2.2 Comics

Comics are a form of visual storytelling that uniquely combine images, text, and narrative progression. They convey meaning not just through individual panels but

through the transitions between them, requiring readers to engage in closure—the cognitive process of inferring unshown actions and connecting discrete events. Speech balloons, thought bubbles, facial expressions, and panel composition all contribute to the story, often in ways that are subtle, ambiguous, or culturally contextual.

From a computational perspective, understanding comics poses several challenges: (1) interpreting visual scenes with high stylistic variance; (2) tracking characters across panels despite changing poses or abstract depictions; (3) aligning dialogue with speaker identities; and (4) understanding temporal and causal relationships in sequential art. Unlike natural images, comic panels are often stylized, symbolic, and highly contextual, making them a unique and rigorous testbed for evaluating VLMs' generalization capabilities.

This thesis proposes a benchmark specifically targeting these challenges, offering tasks that probe a model's ability to understand visual storytelling in comics. By doing so, we aim to investigate the boundaries of current VLMs and lay the groundwork for future research in multimodal narrative understanding.

2.2.1 Benchmarks for Comics

Recently, comic benchmarks have shifted from fundamental tasks—such as object detection, speaker identification, and reading order detection [15]—targeting specific models to more comprehensive tasks tailored for MLLMs. *MangaUB* [16] and *MangaVQA* [17], targeting manga—the Japanese comic art form—assess both panel-level recognition and multi-panel comprehension through manually curated question–answer pairs spanning diverse narrative scenarios. In contrast, *ComicsPAP* [18] and *StripCipher* [19] emphasize multi-panel reasoning tasks such as panel prediction and reordering, highlighting a significant performance gap between current MLLMss and human capabilities in understanding sequential comic narratives.

Table 2.2: Features and statistical information of AI4VA-FG and prior related benchmarks. #QA refer to the number of question-answer pairs. "Public" indicates whether the images are sourced from the public domain, ensuring that the benchmark can be distributed without legal restrictions. AI4VA-FG is, to our knowledge, the only publicly available benchmark for comic understanding that has been systematically evaluated using both SFT and RL.

Benchmark	Task Categories	#QA	SFT	RL	Public
MangaUB	Recognition, Comprehension, Reordering	6,585	×	×	×
StripCipher	Comprehension, Reordering	2,170	✓	×	×
ComicsPAP	Reordering	103,933	✓	×	✓
MangaVQA	Recognition, Comprehension	40,363	✓	×	×
AI4VA-FG	Recognition, Comprehension, Reordering	16,264	✓	✓	✓

2.3 Vision-Language Models (VLMs) for Comics

Traditionally, comic understanding in the AI community has relied on task-specific models, each designed to handle a narrow function such as panel and character detection, reading order prediction, dialogue OCR, captioning, etc. While effective in isolation, these modular approaches lack generalization and often require complex pipelines.

Recent works, such as the Magi series [20, 4], typically employ a multi-stage pipeline architecture, wherein each stage is managed by a task-specific model trained for a distinct subcomponent of the overall system (e.g., panel parsing, captioning, or narrative generation). While this modular approach has demonstrated promising results, its overall effectiveness is inherently constrained by the scope and quality of the training data available for each specialized component. In particular, such pipelines often struggle to generalize in complex scenarios involving numerous characters or intricate visual narratives. This limitation motivates the investigation into applying end-to-end large VLMs for holistic comic understanding, aiming to bypass the dependency on highly curated task-specific modules.

2.4 VLM Post-Training

The post-training phase is essential for enhancing pretrained MLLMs' capabilities for real-world deployment. This stage predominantly encompasses two methodologies: supervised fine-tuning (SFT) and reinforcement learning (RL) [21]. Recently, DeepSeek-R1 [22] demonstrated substantial improvements in text-based reasoning through RL with rule-based rewards, and subsequent studies [23] have further validated the effectiveness of pure RL in enhancing visual reasoning capabilities. Notably, RL demonstrates substantially stronger generalization capabilities than SFT when handling out-of-distribution (OOD) multimodal tasks [24, 21, 25].

Table 2.3: Comparison between SFT and RL.

Aspect	**SFT (Supervised Fine-Tuning)**	**RL (Reinforcement Learning)**
Goal	Imitate reference data	Optimize reward signals
Data	Labeled prompt–response pairs	Prompts + reward signals
Costs	Low for training but high for data	High for training
Complexity	Simple, stable	Complex, instable
Generalization	Low generalizability, replaces existing capabilities	High generalizability, incentivize internal capabilities

2.4.1 Supervised Fine-Tuning (SFT)

The standard objective function for SFT is the Cross-Entropy (CE) Loss, which minimizes the negative log-likelihood of the target sequence given the input prompt.

This encourages the model to predict the exact sequence of tokens present in the training data:

$$\mathcal{L}_{SFT}(\theta) = -\sum_{i=1}^{L} \log P(y_i|y_{<i}, x; \theta)$$

Where:

- θ: Parameters of the current model.
- L: The length of the target output sequence.
- y_i: The i-th token in the target output sequence.
- $y_{<i}$: The sequence of tokens preceding y_i in the target output.
- x: The input prompt.
- $P(y_i|y_{<i}, x; \theta)$: The probability of predicting token y_i given the input x and previous tokens $y_{<i}$, according to the model with parameters θ. This probability is typically derived from the softmax output over the model's vocabulary.

A primary limitation of SFT is its reliance on ground-truth target outputs, which often require extensive manual annotation and may not be scalable for complex or diverse tasks. In our benchmark, the multi-choice questions have ground-truth answer labels, so SFT using questions and answers is possible. Nevertheless, our primary goal is to enhance the general reasoning capabilities of VLMs, enabling them to generalize to a broader range of comic-related tasks. Given the difficulty and cost of generating or annotating reasoning sequences, we turn to RL, the second primary post-training paradigm.

2.4.2 Reinforcement Learning (RL) for VLMs with Verifiable Rewards (RLVR)

RL offers a transformative paradigm for enhancing LLMs for complex downstream tasks. Unlike SFT that relies on static and fully labeled supervision data, RL enables models to learn through trial and error, guided by dynamic and often non-differentiable feedback signals. A widely adopted method in this context is the Proximal Policy Optimization (PPO) algorithm [26], which has been effectively used to align LLMs with human preferences [27].

A central technique driving recent progress is Reinforcement Learning with Verifiable Rewards (RLVR) [28, 22], which applies RL to LLMs or VLMs using task-specific reward signals that can be automatically verified. Typically, RLVR defines the reward based on the correctness of the model's final output—such as assigning a binary reward of 1 or 0 depending on whether the predicted solution to a problem matches the ground truth—thus enabling scalable and objective supervision without

requiring human annotations.

2.4.3 Group Relative Policy Optimization (GRPO)

Group Relative Policy Optimization (GRPO) [29, 22] is an RL algorithm that enhances LLM reasoning, notably in models like DeepSeek-R1, for tasks such as mathematics and coding. It distinguishes itself from PPO [26] by forgoing an explicit value function, instead using the average reward of a group of sampled outputs as a baseline for advantage estimation.

Given an input x from data distribution $\mathcal{D}$, an existing policy $\pi_{\theta_{old}}$, and a reference policy $\pi_{\theta_{ref}}$, GRPO optimizes the policy π_θ by maximizing the following objective :

$$\mathcal{J}_{GRPO}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\mathbb{E}_{(x,a_i)\sim\pi_{\theta_{old}}}\left[\min\left(\frac{\pi_\theta(a_i|x)}{\pi_{\theta_{old}}(a_i|x)}\tilde{r}_i, \text{clip}\left(\frac{\pi_\theta(a_i|x)}{\pi_{\theta_{old}}(a_i|x)}, 1-\epsilon, 1+\epsilon\right)\tilde{r}_i\right)\right] - \beta D_{KL}(\pi_\theta||\pi_{ref})$$

Where:

- θ: Parameters of the current model (i.e., the policy in RL, denoted as π_θ).
- $J(\theta)$: The objective function that GRPO aims to maximize in order to optimize the LLM policy π_θ.
- G: The number of sample responses (i.e. actions) generated for a single input x.
- x: The input prompt.
- a_i: The i-th action (i.e., the response sequence of tokens) sampled by the policy.
- $\pi_\theta(a_i|x)$: The probability of generating action a_i given input x under the current policy π_θ.
- $\pi_{\theta_{old}}(a_i|x)$: The probability of generating action a_i given input x under the previous policy $\pi_{\theta_{old}}$, used to compute the policy ratio.
- $\tilde{r}_i$: The normalized advantage score for the i-th response. This is typically calculated as the difference between the reward of the i-th response and the average reward of the sample group, i.e., $\tilde{r}_i = r_i - \bar{r}$.
- ϵ: The hyperparameter that defines the clipping range $[1-\epsilon, 1+\epsilon]$ to prevent the policy from making large updates.
- β: The coefficient that controls the weight of the KL divergence penalty.

- $D_{KL}(\pi_\theta||\pi_{ref})$: The Kullback-Leibler divergence between the current policy π_θ and a reference policy π_{ref} (often the base model or the model finetuned via SFT), which serves as a penalty for policy drift.

GRPO has been shown to enhance the general visual reasoning capabilities of vision-language models (VLMs) on fundamental tasks such as object counting [24]. Building on this, we investigate its application for improving VLM performance in the more complex domain of comic understanding.

2.4.4 Tool Integrated Reasoning & Thinking with Images

Tool-integrated learning is a paradigm where models are enhanced with access to external tools—such as search engines, Python interpreter, or APIs—to perform tasks that go beyond their internal knowledge. This integration addresses LLMs' intrinsic limitations, which include their inability to fetch real-time data, interact with external systems, or take actions autonomously.

Agentic reinforcement learning (Agentic RL) provides the learning framework for MLLMs to intelligently select and utilize tools. RL enables agents to learn optimal strategies for leveraging external tools and interacting with complex environments through a continuous cycle of trial and error and feedback [30]. This training promotes deeper reasoning and more effective tool use, leading to higher-quality solutions. Reward signals can be designed to encourage successful tool execution and structured outputs.

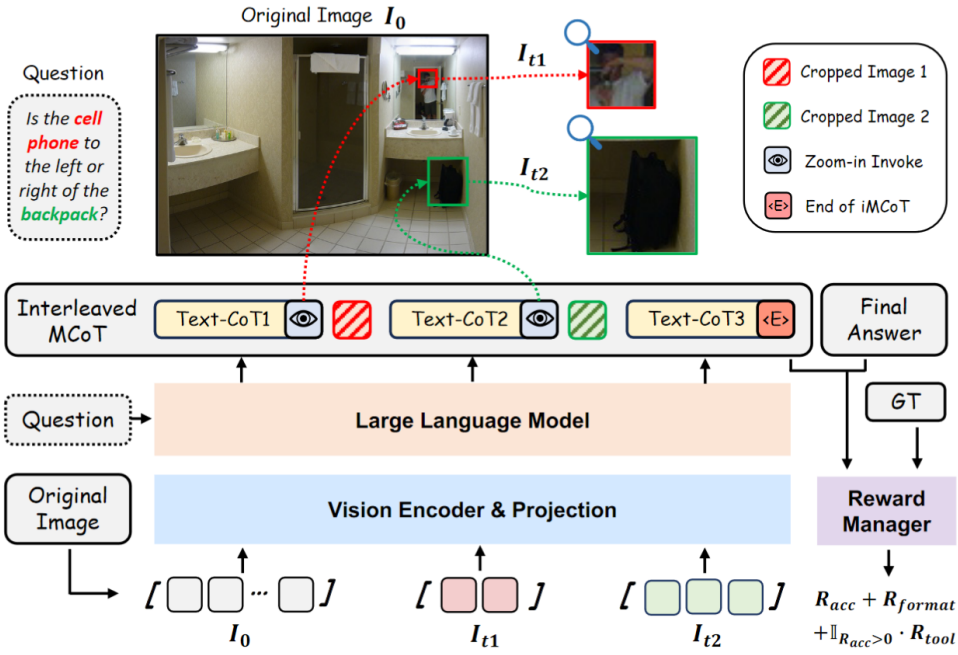

Figure 2.2: The DeepEyes framework [2] leverages agentic RL to train VLMs to progressively crop images on regions of interest through multi-turn interactions.

"Thinking with images" is a recent visual reasoning paradigm in which image manipulation tools—such as zoom-in or cropping—are used to transform the input image, enabling VLMs to better comprehend and reason about visual content. To emulate the human ability to process complex visual information through selective attention, MLLMs can learn to dynamically identify salient regions within an image and adaptively "zoom in" to form a visual chain of thought (CoT) [31].

Recently, *DeepEyes* [2] adopts an end-to-end RL training paradigm (without SFT cold-start) to incentivize tool-assisted, image-based reasoning. Meanwhile, some studies [32, 33] adopt a two-stage post-training approach: firstly, grounding capabilities are established through instruction tuning; subsequently, visual reasoning is enhanced via RL. While these methods predicting variable-size bounding boxes for zoom-in, [34] predicts points and does fixed-size crops to reduce end-to-end RL training costs.

3

AI4VA-FG: A Benchmark for Comics Understanding

3.1 Existing Benchmarks

Existing benchmarks for comics understanding typically collects subtasks of three main categories: *recognition*, *comprehension*, and *reordering*. Recognition tasks focus on identifying visual or textual elements, such as characters, speech balloons, and panels; comprehension tasks aim to evaluate a model's ability to infer narrative context, character intentions, or emotions; reordering tasks evaluate a model's narrative understanding by requiring it to arrange shuffled panels or dialog, or to predict subsequent panels, in a coherent sequence.

- *StripCipher* [19] and *ComicsPAP* [18] employ several variants of reordering tasks to evaluate a model's ability to understand high-level temporal relationships across comic panels. However, these benchmarks do not place emphasis on fine-grained, entity-level recognition and understanding, such as tracking characters across a sequence of panels.

- *MangaUB* [16] and *MangaVQA* [17] include low-level visual question answering tasks but do not provide explicit annotations for regions of interest within the images. This lack of spatial annotations restricts the potential for fine-grained visual grounding and reasoning, as models are unable to directly associate specific image regions with relevant textual descriptions or queries.

- Many existing comic-based benchmarks are constructed using copyrighted datasets such as Digital Comic Museum (DCM) [35] and Manga109 [36]. Due to copyright protections, accessing these datasets typically requires obtaining formal permission from the copyright holders, which limits their accessibility for research and reduces their potential for widespread adoption as standardized evaluation resources.

- Although several prior works have explored the use of supervised fine-tuning (SFT) to adapt models for comic-related tasks, none have investigated reinforcement learning (RL) methods. Given the increasing evidence that RL can significantly improve visual reasoning capabilities [24, 23], it may be par-

ticularly advantageous for challenging tasks that require complex reasoning over multiple panels, such as determining the correct temporal order in panel reordering tasks.

3.2 AI4VA-FG: A Fine-Grained Benchmark Targeting Visual and Narrative Understanding in Comics

Considering the aforementioned limitations of existing benchmarks, we introduce a new comics benchmark, AI4VA-FG (AI4VA Fine-Grained), the first benchmark specifically designed for both low-level and high-level tasks in comics, incorporating entity-level recognition as well as temporal reasoning questions. All imagery is sourced from the public domain, and both the annotations and code are released under a permissive license to support future research.

We develop our benchmark based on the AI4VA dataset [37], which offers a rich and diverse collection of comic-style imagery sourced from two mid-twentieth-century Franco-Belgian comics series, *Placid et Muzo* and *Yves le loup – Bandes Dessinées*, whose faded colors, halftone shading, and hand-lettered typography differ markedly from the natural images and contemporary digital art that predominate in VLM benchmarking. The dataset offers dense annotations of semantic segmentation, ordinal depth, and visual saliency for each comic page, thereby providing a strong foundation for generating visual question answering (VQA) samples tailored to the evaluation of VLMs.

We initially employ a scripted pipeline to generate questions from the segmentation labels provided in AI4VA. These automatically generated questions are then refined through a manual filtering process to ensure clarity and semantic alignment with the visual content. All VQA instances of AI4VA-FG are equipped with bounding box annotations, a design choice that ensures suitability for agentic RL and facilitates the development and evaluation of models across a wide range of visual reasoning capabilities.

3.2.1 Task Definitions

Based on their contextual scope, all questions can be categorized into two distinct types: *single-panel* and *multi-panel*. Single-panel questions are grounded within the content of a single comic panel, while multi-panel questions require reasoning across multiple panels within a page. Each of these two types is further divided into two subcategories: *recognition* and and *understanding* tasks. Recognition tasks focus on extracting explicitly presented information from the image, while understanding tasks involve inferring implicit or internal information embedded in the visual and narrative context of the page.

Table 3.1: Summary of benchmark subtasks and their associated statistics. #Ch. denotes the number of answer choices per multi-choice question, and #QA refers to the total number of VQAs for each subtask. The Character Counting task requires numerical answers, while the Panel Understanding task requires open-text responses that rely on LLM-as-a-Judge [3] for verification; all other tasks utilize multiple-choice answers.

Category	Task	Type	#Ch.	#QA
Single-Panel Understanding	Panel Understanding	open-ended	/	7902
Single-Panel Recognition	Action Recognition	multi-choice	4	1669
	Depth Comparison	multi-choice	2	1125
Multi-Panel Understanding	Dialog Reordering	multi-choice	2	1364
	Panel Reordering	multi-choice	2	1392
Multi-Panel Recognition	Character Identification	multi-choice	4	2368
	Character Counting	numerical	/	444
Total				**16264**

Our benchmark encompasses seven tasks in total, generally arranged in order of increasing difficulty:

- **Panel Understanding**: Evaluates the model's capacity to locate a specific panel within the page layout and comprehend both its visual and textual content.
- **Action Recognition**: Assesses how well the model can recognize a marked character and infer its posture or action based on visual cues.
- **Depth Comparison**: Tests the model's skill in reasoning about spatial relationships by comparing the relative depth of characters and objects within a panel.
- **Dialog Reordering**: Assesses the model's grasp of narrative coherence by requiring it to reconstruct the correct reading order of shuffled dialog balloons. Given that the extracted dialog text is provided in the prompt, this task is generally easier than Panel Reordering.
- **Panel Reordering**: Tests the model's understanding of story progression and visual continuity by asking it to insert a given panel into the correct position among a set of missing panels. Together with Dialog Reordering, this task assesses the model's capability to comprehend and narrate the story flow within a comic sequence.
- **Character Identification**: Evaluates the model's accuracy in determining whether two characters shown in different panels refer to the same entity. This task is motivated by the observation that VLMs often confuse character identities across different panels when captioning an entire page.

- **Character Counting**: Measures the model's accuracy in counting how many times a specific character appears across all panels on the page. Together with Character Identification, these two evaluate the model's effectiveness in tracking characters across multiple panels at both simple and challenging levels of difficulty.

As illustrated in Fig. A.1, for Panel Understanding tasks the panel positions are provided only as textual descriptions rather than visual markings, requiring models to perform grounding solely from the text prompts; in contrast, other tasks explicitly mark relevant characters or panels in the image, making grounding easier.

Figure 3.1: AI4VA-FG's VQA instances for each task. Blue solid-line boxes and orange filled boxes denote real marks appearing in the comic page image, while green dashed boxes indicate implicit regions described in the text prompt.

3.2.2 Dataset Construction

Panel Understanding. The process of constructing this task subset consists of four steps: (1) crop each comic page into individual panels and use Gemini-2.5-flash to generate captions for all panels; (2) based on the panel images and captions,

prompt Gemini-2.5-flash to generate several pairs of original questions and answers; (3) We employ a specialized panel-ordering model to index all panels within each page and generate textual descriptions of their positions, followed by manual verification to ensure positional accuracy; (4) concatenate the positional descriptions with the original questions to form the final VQA triplets *(comic page image, question, answer)*. Since the ground-truth answers for the Panel Understanding task are open-text, an LLM-as-a-Judge [3] approach is employed to verify model responses.

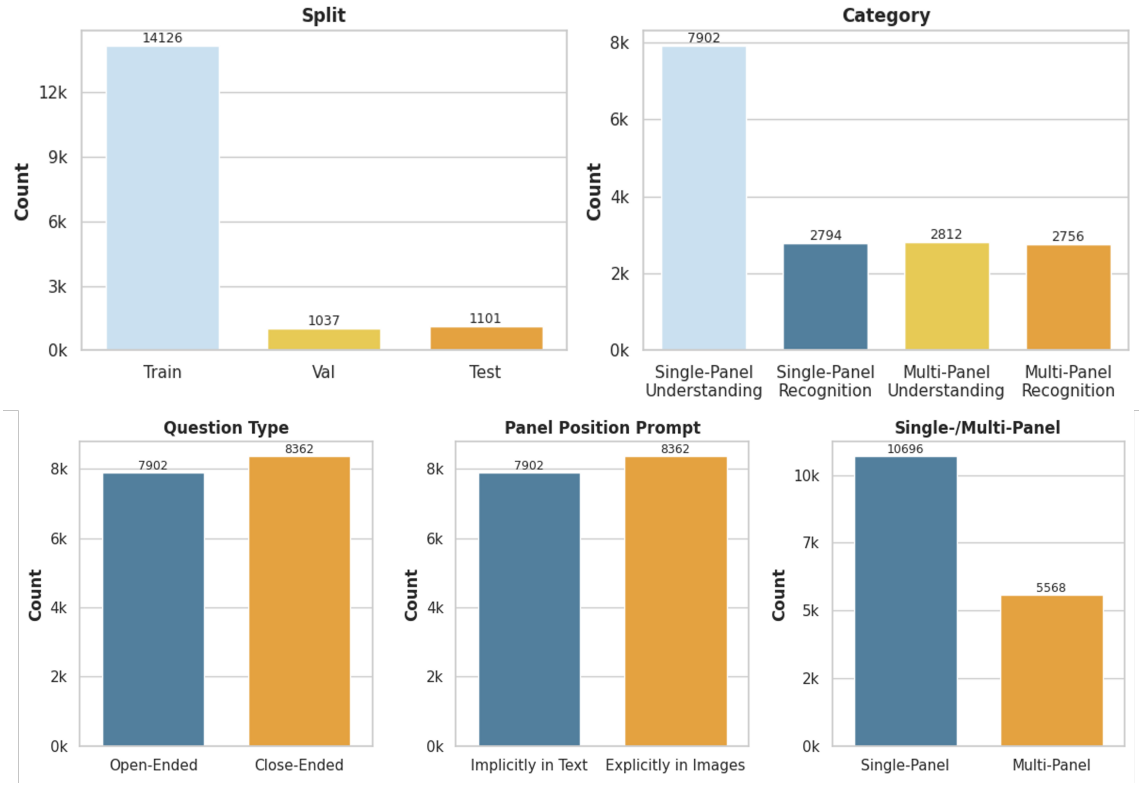

Figure 3.2: Statistics of our AI4VA-FG benchmark.

Closed-Ended Tasks. For all other closed-ended tasks, we develop a pipeline framework to transform segmentation annotations into a QA format compatible with LLMs. Using this framework, we convert AI4VA's segmentation annotations into six tasks comprising roughly 8k triples, and we will release the pipeline to enable the generation of additional VQAs for other comic datasets when large-scale training is required.

All tasks are divided into standard training, validation, and test splits. With the exception of the Dialog and Panel Reordering tasks, which lack defined relevant panels, all other tasks' VQA triples are annotated with bounding boxes of the corresponding panels or characters.

3.3 Performance Analysis

We evaluate selected VLMs on the benchmark test split, with results reported in Table 3.2. Overall, proprietary models outperform their open-source counterparts, and Gemini-2.5-pro achieves the highest accuracy in 6 out of 7 tasks. Interestingly, open-source models that rank higher on general VLM leaderboards still fail to surpass commercial models on our benchmark. Despite these advances, a substantial gap persists between current models and human-level comic understanding, primarily due to pronounced limitations in spatial perception, character tracking, and multi-panel narrative construction.

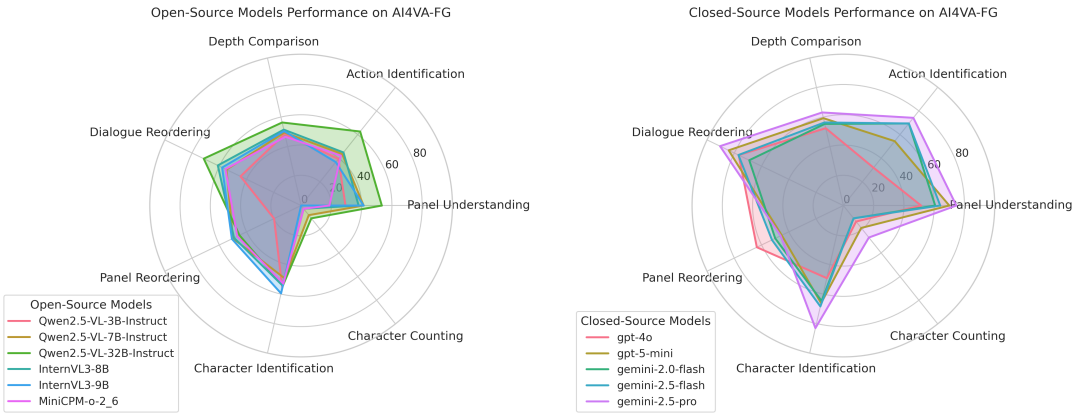

Figure 3.3: Open-source models v.s. closed-models on AI4VA-FG across tasks. Gemini-2.5-pro achieves the highest accuracy in 6 out of 7 tasks.

Depth Comparison. Surprisingly, the majority of evaluated models exhibit poor spatial perception when processing comic images, performing near random when comparing entity depth in the image. Although spatial reasoning is a fundamental capability for VLMs and has been included in several VQA benchmarks [38, 39], it remains a known weakness of current general-purpose VLMs. Compared to depth perception in real-world images [38], GPT-4o's performance on comics is notably more random, suggesting that the stylistic nature of comic drawings introduces additional challenges for spatial reasoning.

Dialog & Panel Reordering. While Gemini-2.5-flash and GPT-5-mini achieve promising results on the Dialog Reordering tasks, their performance on Panel Reordering remains close to random selection (50%). In Dialog Reordering, OCR-extracted dialog text is supplied within the prompt, aiding the model in reconstructing the narrative flow across adjacent panels. However, in the absence of such textual guidance, VLMs exhibit markedly constrained ability to compose coherent narratives solely from discrete sequential images.

Character Identification & Counting. Each image in AI4VA contains on average 13 panels, making it particularly challenging to track all appearances of a

[1]Gemini's high accuracy on Panel Understanding cannot be regarded as a fair measure, since both the questions and answers in this task were generated by Gemini.

character across panels, especially when individual appearances are too small to be reliably identified. This difficulty accounts for the poor performance of all models on the Character Counting task. The Character Identification task can be viewed as a simplified variant of Character Counting, as the model is only required to identify two marked characters rather than enumerate all appearances. While commercial models achieve promising accuracy on identifying characters, the evaluated open-source VLMs perform at near-random levels.

Table 3.3: Accuracy (%) of selected models on AI4VA-FG with entire-page and individual-panel inputs, respectively. Zoom-in on relevant individual panels yields significantly improved accuracy.

Model	Panel Understanding	Action Recognition	Depth Comparison	Character Identification
Qwen2.5-VL-7B	41.33	43.54	49.51	49.26
(zoom-in)	+17.34 58.67	+12.24 55.78	+4.83 54.34	+3.68 52.94
GPT-4o	51.33	31.97	52.43	49.22
(zoom-in)	+21.34 72.67	+23.13 55.10	+6.79 59.22	+14.75 63.97
Gemini-2.5-Flash	64.00	69.39	56.31	68.38
(zoom-in)	+16.00 80.00	+6.12 75.51	+22.33 78.64	+6.62 75.00

Does Zooming-In Improve Performance? For humans, these tasks are challenging when viewing the entire page at a glance, but become considerably easier when focusing on the relevant panels. This motivates us to further compare the performance of entire-page versus individual-panel inputs for the single-panel tasks. We observe accuracy improvements across selected tasks when zooming solely into the relevant panels, likely due to the removal of unrelated content. Notably, zooming in yields a substantial 22.33% improvement on the challenging Depth Comparison task for Gemini-2.5-flash. This finding is consistent with [40], who argue that while VLMs can attend to fine-grained visual details in high-resolution images, their perceptual capacity is constrained by the spatial extent of the input. These results highlight the importance of incorporating zoom-in mechanisms that enable models to focus on salient, detailed regions of a comic page, such as individual panels or characters.

Table 3.2: Evaluation results of state-of-the-art VLMs on AI4VA-FG. While proprietary models generally outperform open-source counterparts, a performance gap remains to human-level understanding. Notably, most models perform close to random chance on Depth Comparison, Panel Reordering, and Character Counting. Best and second-best performance values (that exceed random-guess accuracy) are indicated using **bold** and underlined formatting, respectively.

Model	Panel Understanding	Action Recognition	Depth Comparison
Random	/	25.00	50.00
Qwen2.5-VL-3B-Instruct	29.33	39.46	48.54
Qwen2.5-VL-7B-Instruct	41.33	43.54	49.51
Qwen2.5-VL-32B-Instruct	53.33	62.59	56.31
InternVL3-8B	38.00	44.90	51.46
InternVL3-9B	41.33	36.73	50.49
MiniCPM-o-2.6 (8B)	18.67	42.18	46.60
GPT-4o-2024-08-06	51.33	31.97	52.43
GPT-5-mini-2025-08-07	<u>70.00</u>	54.42	<u>59.22</u>
Gemini-2.0-Flash	60.67[1]	<u>69.39</u>	55.34
Gemini-2.5-Flash	64.00[1]	<u>69.39</u>	56.31
Gemini-2.5-pro	**74.67**[1]	**74.15**	**63.11**

Model	Dialog Reordering	Panel Reordering	Character Identification	Character Counting
Random	50.00	50.00	50.00	/
Qwen2.5-VL-3B-Instruct	44.44	19.84	55.88	2.70
Qwen2.5-VL-7B-Instruct	54.76	50.00	49.26	8.11
Qwen2.5-VL-32B-Instruct	71.43	45.24	54.41	10.81
InternVL3-8B	61.11	49.21	54.41	2.70
InternVL3-9B	57.94	50.79	59.56	0.00
MiniCPM-o-2.6 (8B)	55.56	46.83	52.94	2.70
GPT-4o-2024-08-06	76.80	**63.49**	49.22	13.51
GPT-5-mini-2025-08-07	<u>84.13</u>	45.24	66.18	<u>18.92</u>
Gemini-2.0-Flash	69.05	50.00	64.71	10.81
Gemini-2.5-Flash	76.98	52.38	<u>68.38</u>	10.81
Gemini-2.5-pro	**90.48**	46.03	**83.09**	**27.03**

4

Methodology

To improve the comic understanding capabilities of VLMs and narrow the gap between open-source models, proprietary systems, and humans, we fine-tune Qwen2.5-VL-7B-Instruct using both SFT and RL. For SFT, we explore (i) simply fine-tuning with question–answer pairs (denoted as SFT-S) and (ii) distillation with correct reasoning trajectories generated by Gemini-2.5-flash (denoted as SFT-R). For RL, we experiment with both vanilla RL and **Region-Aware RL**, the latter designed to incentivize zoom-in operations on relevant image regions explicit guidance for tool-usage accuracy.

4.1 Enable "Thinking with Images" via Agentic Reinforcement Learning

We adopt a two-stage Agentic RL framework: (1) a brief **warm-start phase** that leverages only basic tool usage rewards to establish tool-calling behavior, and (2) a main **RL training phase** that incorporates the complete reward structure to incentivize accurate and effective zoom-in actions. In contrast to other two-stage approaches that rely on a SFT cold-start, our warm-start phase remains entirely within the RL paradigm, differing solely in the reward configuration. As a result, it does not require any curated SFT datasets consisting of manually synthesized tool-calling trajectories.

Reward Strategy. In the context of VLMs, outcome-based rewards play a key role in steering models toward effective reasoning and decision-making. In our RL training phase, the total reward structure consists of three parts: an accuracy reward R_{acc}, a formatting reward R_{format}, and a **tool usage reward** R_{tool}. The accuracy reward measures whether the final answer is correct, while the formatting reward penalizes improperly structured outputs. The tool usage reward is given when an external tool is called correctly during the reasoning process. Formally, for a reasoning trajectory τ, the total reward is:

$$R(\tau) = R_{\text{format}}(\tau) + R_{\text{acc}}(\tau) + R_{\text{tool}}(\tau), \tag{4.1}$$

The tool usage reward depends both on whether the external tool is invoked appropriately and on the accuracy of the tool's output relative to the given question.

DeepEyes [2] employs a strategy in which a constant tool usage bonus is added to the total reward only when the final answer is correct. However, in our setting, since each question includes a ground-truth region of interest (e.g., a character or panel region on the page), we propose a variant of the tool usage reward that more effectively encourages correct tool usage by explicitly measuring spatial accuracy:

$$R_{\text{tool}}(\tau) = (1 + \mathbb{I}_{R_{\text{acc}}(\tau)>0})(R_{\text{tool-count}}(\tau) + R_{\text{tool-acc}}(\tau)) \tag{4.2}$$

where $\mathbb{I}$ is an indicator function that activates an additional tool-usage bonus only when the final answer is correct, ensuring that the tool usage likely contributes to the outcome. $R_{\text{tool-count}}(\tau)$ denotes the reward component evaluating whether the number of zoom-in tool invocations in the reasoning trajectory matches the expected count, and $R_{\text{tool-acc}}(\tau)$ represents the accuracy-based bonus awarded for correct tool usages:

$$R_{\text{tool-acc}}(\tau) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \text{IoU}(\tau_i) \tag{4.3}$$

Here, τ_i denotes the sub-trajectory corresponding to the i-th tool usage, and m is the number of zoom-in tool invocations. For each predicted zoom-in bounding box, the Intersection over Union (IoU) is computed between the predicted box and its corresponding target region in the image. The accuracy bonuses $R_{\text{tool-acc}}(\tau_i)$ are summed to give higher rewards when multiple zoom-in operations are correctly performed. The normalization by $\sqrt{m}$ stabilizes the reward distribution when multiple bounding boxes are output for tasks such as Character Identification & Counting.

4.2 Experiment Setups

We train Qwen2.5-VL-7B-Instruct on 8 × H800 (80G) GPUs, using LLaMA-Factory[1] and verl[2] frameworks for SFT and RL respectively. We adopt GRPO [29] algorithm for RL training. See Appendix A.1 for details of the hyperparameter configurations.

We train the open-ended and closed-ended tasks in Tab. 3.1 independently and report results in Tab. 5.1. For closed-ended tasks, SFT is trained jointly across all tasks, while RL first uses sequential training across three categories followed by merged training, since starting with all tasks simultaneously causes convergence instability. Under sequential training, earlier tasks exhibit performance degradation as new categories are introduced, suggesting that the 7B base model, combined with the current dataset size, lacks sufficient capacity to generalize across all tasks concurrently. We also conduct category-wise training for both SFT and RL, which leads to greater improvements on most tasks, and confirm that the overall interpretation of results in Chapter 5 remains consistent.

[1] https://github.com/hiyouga/LLaMA-Factory
[2] https://github.com/volcengine/verl

5

Experiment

5.1 Main Results

Table 5.1: Accuracy (%) of finetuned models on AI4VA-FG tasks.

Model	Panel Understanding	Action Recognition	Depth Comparison
Qwen2.5-VL-7B	41.33	43.54	49.51
SFT-S (vanilla)	+8.67 50.00	+26.53 70.07	-3.88 45.63
SFT-R (reasoning)	+7.34 48.67	+24.49 68.03	-3.88 45.63
RL (vanilla)	+14.00 **55.33**	+28.57 <u>72.11</u>	+2.92 <u>52.43</u>
RL (Region-Aware)	+10.00 <u>51.33</u>	+32.64 **76.19**	+7.28 **57.28**

Model	Dialog Reordering	Panel Reordering	Character Identification	Character Counting
Qwen2.5-VL-7B	54.76	50.00	49.26	8.11
SFT-S (vanilla)	+0.80 55.56	-2.98 47.02	-0.04 49.22	+10.81 **18.92**
SFT-R (reasoning)	+14.29 **69.05**	-0.79 49.21	+14.71 63.97	+2.70 <u>10.81</u>
RL (vanilla)	+5.56 <u>60.32</u>	-2.38 47.62	+16.92 <u>66.18</u>	-2.70 5.41
RL (Region-Aware)	/	/	+22.06 **71.32**	-5.41 2.70

SFT v.s. RL. On most tasks, both SFT and RL yield significant improvements. Among the two SFT settings, SFT-R consistently outperforms SFT-S, demonstrating that CoT distillation enhances visual reasoning even for low-level recognition tasks. Furthermore, except for Dialogue Reordering and Character Counting, RL outperforms SFT and matches or exceeds proprietary models on certain tasks.

However, on the two ordering tasks, RL brings very limited improvement and lags far behind distillation with Gemini's CoT trajectories. This may be because the internal reordering capability of the 7B base model is substantially weaker than that of Gemini-2.5, and GRPO can only amplify existing abilities but struggles to create entirely new ones [25]. Specifically, we also apply RL on top of the distilled

model, but its performance surprisingly gradually degrades: the model forgets the ordering ability inherited from Gemini-2.5-flash and fails to acquire new effective reasoning patterns during RL training.

Region-Aware RL. Region-Aware RL optimizes two objectives: grounding IoU and VQA accuracy. Since ground-truth bounding boxes are unavailable for the two reordering tasks, we fine-tune Qwen-2.5-VL-7B on the remaining five tasks using Region-Aware RL. The results suggest that the model possesses a strong grounding ability, reaching nearly 80% IoU from 20% for zoom-in operations if trained on Action Recognition and Depth Comparison only. Grounding accuracy on the Panel Understanding task is lower, as the relevant panel is specified only in the prompt but not explicitly marked in the image, which increases the likelihood of errors when the model attempts to localize the correct panel. Notably, when multiple zoom-in operations are performed, the second operation is less accurate than the first, suggesting that limited context length constrains the model's grounding accuracy.

Tab. 5.1 demonstrates Region-Aware RL outperforms vanilla RL and even surpasses Gemini-2.5-Flash on three recognition tasks, achieving performance comparable to the latter with manual cropping (see Table.3.3). Its weaker improvement on Panel Understanding may stem from imprecise panel localization in this task. This result highlights that a smaller model, when equipped with appropriate post-training strategies, can exceed the performance of a much larger model on specific tasks. It also underscores the potential of tool-augmented reasoning to enhance model performance in scenarios involving large and visually dense contexts.

Table 5.2: Zoom-in Statistics of models finetuned via region-aware RL.

Task	Panel Understanding	Action Recognition	Depth Comparison	Character Identification
Avg. #Toolcall	1.01	1.10	0.93	1.88
Avg. IoU	0.565	0.862	0.847	0.835 (1^{st}: 0.842; 2^{nd}: 0.829)

Notably, Region-Aware RL yields no improvement on the Character Counting task. Ideally, if a model can sequentially zoom into each panel, it should discover all occurrences of a target character; but in practice, the model fails to learn a reliable "crop-all-panels" behavior via pure RL. We attribute this failure to two factors: (i) limited training samples for this task when jointly trained with other tasks, and (ii) degraded grounding accuracy as context length increases—when more than two panels have already been cropped. Addressing these issues likely requires stronger supervision or curriculum strategies to stabilize sequential zooming and bounding-box prediction.

Figure 5.1: An example multi-turn conversation with zoom-in tool calling.

5.2 Ablations

In-Domain, Cross-Task Generalization. To investigate the generalization ability of the post-training methods, we train the model on each of the 3 closed-ended categories and evaluate the finetuned model on the other two. We observe that RL exhibits certain in-domain generalizability as [21, 24] argues. For instance, Action Recognition benefits most from RL: training on multi-panel tasks also brings improvements on Action Recognition without obvious degradation on other tasks.

To assess whether this gain results merely from additional training, we performed an ablation study in which the model was trained with RL using only format rewards and random accuracy rewards. No appreciable performance gains were observed, suggesting that the improvement is driven by the model's ability to leverage meaningful in-domain reward signals rather than by increased training alone. Nevertheless, no performance gains are observed on Depth Comparison or Panel Reordering, supporting the argument that GRPO primarily amplifies existing capabilities already acquired during pre-training rather than enabling new forms of reasoning [25].

Interestingly, SFT does not solely overfit to the supervised task, but can also demonstrate cross-task generalization: fine-tuning the base model with reasoning trajectories generated by Gemini-2.5-flash only for the reordering tasks results in superior

performance (69.85%) also on Character Identification, even surpassing the model trained on this task via RL (69.12%). While Gemini-2.5-flash achieves high accuracy and generates high-quality reasoning on the Dialogue Reordering task, this capability can be transferred to other tasks via distillation. Nevertheless, the extent of SFT in-domain generalization is inherently constrained by the amount and quality of the supervision data.

Table 5.3: Cross-Task Generalization Performance. "SFT-R (character)" refers to the model finetuned on Character Identification & Counting tasks via SFT-R, while "RL (reorder)" refers to the model finetuned on Dialog & Panel Reordering tasks via vanilla RL. RL demonstrates stronger in-domain cross-task generalization than SFT, particularly for recognition-oriented tasks.

Model	Panel Understanding	Action Recognition	Depth Comparison	Dialog Reordering	Panel Reordering	Character Identification	Character Counting
Qwen2.5-VL-7B	41.33	43.54	49.51	55.56	50.00	50.78	8.11
SFT-R (action & depth)	-2.00 39.33	63.95	42.72	-20.64 34.92	-35.71 14.29	-8.13 42.65	-2.70 5.41
SFT-R (reorder)	+10.00 **51.33**	+8.16 51.70	-6.79 42.72	71.43	50.79	+19.07 **69.85**	8.11
SFT-R (character)	+5.34 46.67	43.54	-1.94 47.57	-4.77 50.79	-10.32 39.68	63.28	2.70
RL (action & depth)	+0.67 42.00	76.19	58.25	-2.39 53.17	-0.79 49.21	+10.25 **61.03**	+2.70 10.81
RL (reorder)	-2.00 39.33	+22.45 **65.99**	-6.79 42.72	57.94	44.44	+11.72 **62.50**	-2.70 5.41
RL (character)	+2.00 43.33	+21.09 **64.63**	49.51	-1.59 53.97	+0.79 50.79	69.12	10.81

Cross-Domain Generalization. We further evaluate the fine-tuned model on MangaVQA [17], a benchmark for manga that is closely related to but distinct from comics. Neither SFT nor RL demonstrates cross-domain generalization on this dataset, while SFT leads to significant degradation. This limitation can be attributed to pronounced domain shifts, including differences in artistic style, panel layout, and text density between Western comics and Japanese manga.

Table 5.4: Performance of finetuned models on MangaVQA-test and MMLU-val. RL exhibits minimal degradation on general-domain tasks, whereas SFT incurs higher cross-domain drops.

Method	MangaVQA-test	MMLU-val
Original	22.78	53.78
SFT-R	10.58	51.11
RL	22.00	53.56

In addition, we evaluate the general reasoning capabilities of the RL-trained vision model using the widely adopted MMMU [14] benchmark's validation split . After fine-tuning on six closed-ended tasks, the RL-optimized model exhibits minimal performance degradation on general-domain tasks in MMMU, whereas the SFT-fine-tuned model experiences greater performance drops, indicating that RL may confer better cross-domain robustness than SFT.

Reward Strategy. While DeepEyes [2] argues that end-to-end RL alone is sufficient to enable tool usage, concurrent works [32, 33] highlight the necessity of an

SFT cold-start phase for achieving stable and effective learning. In our experiments, we found that fully end-to-end RL often leads to reward hacking during the early stages of training—where the model exploits only easily achievable components of the reward function, such as format correctness, without improving on more substantive objectives such as tool use or reasoning optimization. To address this, we adopt a two-phase RL strategy, in which the warm-start phase employs a simplified reward formulation to guide the model toward meaningful tool-using behaviors before full reward optimization begins.

We further experimented by removing the constant coefficient in the reward rule Equation 4.2, making the tool usage reward R_{tool} conditional on the correctness of the final answer. This modification led to slow convergence in both tool accuracy and overall task accuracy: the model performs poorly on both targets at the early stage so that conditioning one on the other can greatly slow the learning of tool calling. The results suggest that tool usage should consistently be rewarded whenever zoom-in operations are known to be beneficial, in order to enable more efficient tool learning.

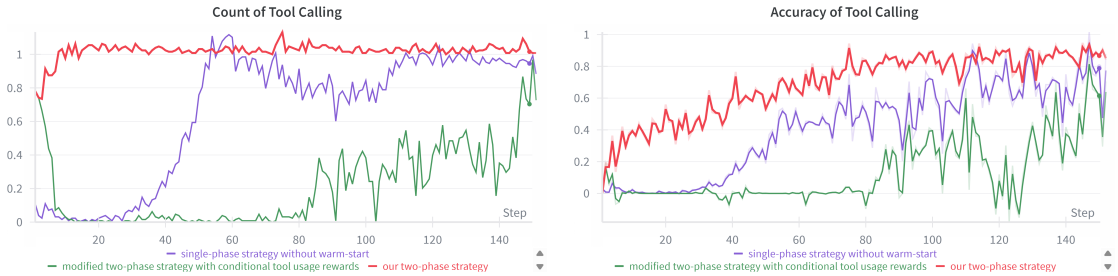

Figure 5.2: Comparison of three reward strategies: (1) single-phase strategy without warm-start, (2) modified two-phase strategy with tool usage reward conditional on the final answer's correctness, and (3) our two-phase strategy (displaying only the second phase starting after 16 warm-start steps). Strategy (1) yields the most efficient tool learning.

6

Conclusion

We presented AI4VA-FG, the first fine-grained benchmark for comic understanding with VLMs, spanning both low-level recognition and high-level reasoning tasks with dense annotations. Through extensive evaluation of state-of-the-art models, we revealed persistent weaknesses in spatial perception, character tracking, and multi-panel narrative construction, underscoring the gap between open-source and proprietary systems.

To mitigate these challenges, we examined post-training strategies, showing that both SFT and RL can yield cross-task generalization, while our proposed Region-Aware RL leverages zoom-in operations to improve grounding and narrative reasoning. Together, our benchmark and methods establish a foundation for advancing multimodal reasoning in the domain of comics.

6.1 Limitations

Despite our significant efforts to construct a comprehensive benchmark and enhance current models' comic understanding capabilities, several limitations remain. First, we do not explore the latest state-of-the-art RL algorithms, which have shown promising results in recent research on VLM post-training. Second, although our benchmark covers a variety of recognition and reasoning tasks, it does not encompass all aspects of comic understanding, leaving room for further expansion. Finally, our focus is primarily on visual comprehension, with less emphasis on visual generation, which may hold greater practical value in real-world applications such as automated comic storytelling.

6.2 License and Copyright

All comic pages in our dataset are sourced from two mid-twentieth-century Franco-Belgian series, *Placid et Muzo* and *Yves le loup – Bandes Dessinées*, both of which are in the public domain, eliminating copyright concerns for our work and future research. We will release all code for dataset preparation, benchmark evaluation, and model training to ensure full reproducibility and to support future studies in comic understanding and multimodal learning.

6.3 Future Work

In future research, more existing comic datasets can be expanded into comprehensive benchmarks for VLMs with our pipeline, enabling large-scale training and more robust evaluation of visual reasoning capabilities.

Beyond recognition and ordering tasks, we can also incorporate additional fundamental tasks in the comic domain, such as speaker identification, to provide a more complete assessment of comic understanding beyond the current recognition and ordering tasks.

Furthermore, the focus can be extended from understanding to generation. For instance, our benchmark could be repurposed to evaluate comic storytelling by VLMs, replacing image inputs with generated captions, thus systematically assessing models' narrative generation and multimodal creative reasoning abilities.

We are also exploring the use of our region-aware RL framework to enable panel-level fine-grained captioning of comic pages. In this setting, the model first zooms into each panel before generating captions, thereby improving its ability to preserve fine-grained details. However, the effectiveness of this approach is constrained by the base model's tool-calling capability under long-context conditions.

Bibliography

[1] Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

[2] 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. 2025. URL https://arxiv.org/abs/2505.14362.

[3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

[4] Ragav Sachdeva and Andrew Zisserman. From panels to prose: Generating literary narratives from comics, 2025. URL https://arxiv.org/abs/2503.23344.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[7] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

[10] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*, 2024.

[12] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

[13] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

[14] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

[15] Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. Comix: A comprehensive benchmark for multi-task comic understanding, 2024. URL https://arxiv.org/abs/2407.03550.

[16] Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. Mangaub: A manga understanding benchmark for large multimodal models, 2024. URL https://arxiv.org/abs/2407.19034.

[17] Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding, 2025. URL https://arxiv.org/abs/2505.20298.

[18] Emanuele Vivoli, Artemis Llabrés, Mohamed Ali Souibgui, Marco Bertini, Ernest Valveny Llobet, and Dimosthenis Karatzas. Comicspap: understand-

ing comic strips by picking the correct panel, 2025. URL https://arxiv.org/abs/2503.08561.

[19] Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Yixin Yang, Qingxiu Dong, Weiyao Luo, Yifan Pu, Yiru Wang, Xiangdi Meng, Wenjie Li, and Zhifang Sui. Beyond single frames: Can lmms comprehend temporal and contextual narratives in image sequences?, 2025. URL https://arxiv.org/abs/2502.13925.

[20] Ragav Sachdeva, Gyungin Shin, and Andrew Zisserman. Tails tell tales: Chapter-wide manga transcriptions with character names, 2024. URL https://arxiv.org/abs/2408.00298.

[21] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL https://arxiv.org/abs/2501.17161.

[22] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[23] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2503.01785.

[24] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02.

[25] Neel Rajani, Aryo Pradipta Gema, Seraphina Goldfarb-Tarrant, and Ivan Titov. Scalpel vs. hammer: Grpo amplifies existing capabilities, sft replaces them, 2025. URL https://arxiv.org/abs/2507.10616.

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[27] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[28] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning, 2024. URL https://arxiv.org/abs/2410.15115.

[29] 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, 2024. URL https://arxiv.org/abs/2402.03300.

[30] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms, 2025. URL https://arxiv.org/abs/2504.11536.

[31] 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, 2024. URL https://arxiv.org/abs/2403.16999.

[32] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning, 2025. URL https://arxiv.org/abs/2505.15966.

[33] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl, 2025. URL https://arxiv.org/abs/2505.15436.

[34] Sunil Kumar, Bowen Zhao, Leo Dirac, and Paulina Varshavskaya. Reinforcing vlms to use tools for detailed visual reasoning under resource constraints, 2025. URL https://arxiv.org/abs/2506.14821.

[35] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. *Journal of Imaging*, 4 (7), 2018. ISSN 2313-433X. doi: 10.3390/jimaging4070089. URL https://www.mdpi.com/2313-433X/4/7/89.

[36] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020. doi: 10.1109/MMUL.2020.2987895.

[37] Peter Grönquist, Deblina Bhattacharjee, Bahar Aydemir, Baran Ozaydin, Tong Zhang, Mathieu Salzmann, and Sabine Süsstrunk. Unlocking comics: The ai4va dataset for visual understanding, 2024. URL https://arxiv.org/abs/2410.20459.

[38] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

[39] Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning, 2025. URL https://arxiv.org/abs/2505.12448.

[40] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms, 2025. URL https://arxiv.org/abs/2502.17422.

A

Appendix

A.1 Experiment Setups for Finetuning

In RL training on closed-ended tasks, the model is first trained sequentially on three categories (up to 200 steps each), and then jointly on all categories for an additional 200 steps.

Table A.1: Hyper-Parameters for RL (GRPO) Training

Hyper-parameter	**Value**
Learning Rate	1×10^{-6}
Number of Steps	200
Rollout Batch Size	16
PPO Mini Batch Size	16
Num of Responses per Sample	8
Max Prompt Length	10280
Max Response Length	4096
Max Response Length (Region-Aware)	4096 * 5
KL Coefficient	0.04
Warmup Ratio	0.0
Rollout Engine	vLLM (0.8.2)
RL Engine	verl (0.2.0.dev0)
Number of GPUs	8

Table A.2: Hyper-Parameters for SFT Training

Hyper-parameter	**Value**
Learning Rate	1×10^{-5}
Number of Epochs	3
Batch Size	16
Optimizer	AdamW
Learning Rate Scheduler	cosine
Warmup Ratio	0.1
Number of GPUs	4

A.2 Prompt Templates

Prompt for captioning each panel and generate QA pairs:

```
"""Describe the content of the comic page in detail, including characters, actions, and any notable elements.
Then, using the given image and the textual information written in it, create {num_qa_pairs} VQA questions.
Format the output in the following way:
Caption: [Caption content]
Question: [Question content]
Answer: [Answer content]
Question: [Question content]
Answer: [Answer content]
...
Avoid subjective questions that could lead to ambiguous interpretations, and instead create questions that can be
objectively answered based on the facts presented in the image. Also, do NOT include OCR-style text recognition questions;
instead, create questions that test understanding of the visual content."""
```

Prompts for questions:

```
(1) Prompt that disables thin:
"""<|image|>
{Question}  Output only the final answer (choice) in <answer> </answer> tags."""

(2) Prompt that encourages thinking:
"""<|image|>
{Question}  Output the thinking process in <think> </think> (if needed) and final answer (choice) in <answer> </answer>
tags."""

(3) Prompt that enables tool calling:
"""<|image|>
{Question}  Think first, call **image_zoom_in_tool** if needed, then answer. Format strictly as: <think>...</think>
<tool_call>...</tool_call> (if tools needed) <think>...</think><answer>...</answer>."""
```

System prompt for enabling tool calling (modified from DeepEyes):

```
"""You are a helpful assistant.

# Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools> XML tags:
<tools>
{"type":"function","function":{"name":"image_zoom_in_tool","description":"Zoom in on a specific region of an image by
cropping it based on a bounding box (bbox) and an optional object label.","parameters":{"type":"object","properties":
{"bbox_2d":{"type":"array","items":{"type":"number"},"minItems":4,"maxItems":4,"description":"The bounding box of the
region to zoom in, as [x1, y1, x2, y2], where (x1, y1) is the top-left corner and (x2, y2) is the bottom-right
corner."},"label":{"type":"string","description":"The name or label of the object in the specified bounding box
(optional)."}},"required":["bbox"]}}}
</tools>

# How to call a tool
Return a json object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>

**Example**:
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [10, 20, 100, 200], "label": "the panel with a boxed character"}}
</tool_call>"""
```

A.3 Example of page-level comic captions generated by VLMs

When prompted to caption a comic page, Gemini-2.5-Flash generated captions for only a subset of all panels and is likely to mix up content across panels.

```
<panel bbox='[52, 235, 126, 403]'>
Panel 1
...
</panel>

...

<panel bbox='[147, 747, 220, 919]'>
Panel 8
### Characters & Actions
The dog is now standing next to the deer, who is on the right side of the panel. The dog is pointing to the deer, and the deer is looking at the dog with a shocked
expression. A small, gray pig is on the left, looking at the dog with a look of astonishment.

### Facial Expressions & Emotions
The dog has a proud and confident expression. The deer has a look of shock and disbelief. The pig has a look of astonishment. The characters' expressions convey a
sense of pride and surprise.

### Setting & Background
The setting is a simple, white background with a brown floor. There are no other background details. The focus is on the characters and their actions.

### Text & Dialogue
The dialogue is in a white speech bubble with a black outline, coming from the dog. The text is in a clear, handwritten font that says, "Attention! Vous n'avez pas
qu'à vous en faire! Il y a trois! Ouf!" which means "Careful! You don't just have to worry! There are three! Phew!". The dog's dialogue is confident and
reassuring, contrasting with the deer's shocked expression.

### Artistic Style & Visual Elements
The art style is consistent. The composition is a medium shot of the characters. The perspective is a straight-on shot. The use of motion lines around the barbell
and the dog's arm suggests movement and effort.

### Panel Transitions
This panel transitions from the previous one by showing the dog's success. The previous panel built anticipation, and this one reveals the dog's success. The shift
from a moment of effort to a moment of triumph changes the story's pacing. The bear's impressed expression provides a direct reaction to the dog's success.
</panel>
<panel bbox='[239, 236, 312, 403]'>
Panel 9
### Characters & Actions
The dog is now standing next to the deer, who is on the right side of the panel. The dog is pointing to the deer, and the deer is looking at the dog with a shocked
expression. A small, gray pig is on the left, looking at the dog with a look of astonishment.

### Facial Expressions & Emotions
The dog has a proud and confident expression. The deer has a look of shock and disbelief. The pig has a look of astonishment. The characters' expressions convey a
sense of pride and surprise.

### Setting & Background
The setting is a simple, white background with a brown floor. There are no other background details. The focus is on the characters and their actions.

### Text & Dialogue
The dialogue is in a white speech bubble with a black outline, coming from the dog. The text is in a clear, handwritten font that says, "Attention! Vous n'avez pas
qu'à vous en faire! Il y a trois! Ouf!" which means "Careful! You don't just have to worry! There are three! Phew!". The dog's dialogue is confident and
reassuring, contrasting with the deer's shocked expression.

### Artistic Style & Visual Elements
The art style is consistent. The composition is a medium shot of the characters. The perspective is a straight-on shot. The use of motion lines around the barbell
and the dog's arm suggests movement and effort.

### Panel Transitions
This panel transitions from the previous one by showing the dog's success. The previous panel built anticipation, and this one reveals the dog's success. The shift
from a moment of effort to a moment of triumph changes the story's pacing. The bear's impressed expression provides a direct reaction to the dog's success.
</panel>
<panel bbox='[239, 403, 312, 574]'>
Panel 10
...
</panel>
```

Figure A.1: Example of page-level comic captions generated by Gemini-2.5-Flash. Only 10 out of 23 panels (see Fig. A.1) were recognized and captioned with some panel contents being incorrectly mixed (e.g., panel 8 and panel 9), and the inter-panel transitions are wrong interpreted.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY