MANGO: Manga Active Narrative Grounding Optimization

Hao Qiu[1]*, Junyan Wang[2*†], Zheyuan Liu[2]
Lei Fan[3], Hong Jia[4], Lianbo Guo[5], Zhulin Tao[1†]
[1]Communication University of China
[2]Australian Institute for Machine Learning, Adelaide University
[3]University of New South Wales [4]University of Auckland
[5]Huazhong University of Science and Technology
qiuhao@cuc.edu.cn junyan.wang&zheyuan.liu@adelaide.edu.au
lei.fan1@unsw.edu.au hong.jia@auckland.ac.nz lbguo@hust.edu.cn
taozl@cuc.edu.cn

Abstract

Manga visual question answering requires models to answer questions over panel-based visual narratives, where relevant evidence is distributed across ordered panels, embedded text, recurring characters, and implicit event transitions. This structure makes passive page encoding insufficient, as the model must identify which panels to inspect, what clues to retain, and when the accumulated evidence is sufficient for answering. We propose MANGO (**M**anga **A**ctive **N**arrative **G**rounding **O**ptimization), an unsupervised framework for active manga visual question answering. MANGO introduces Active Narrative Sketching (ANS), which iteratively selects panels, extracts concise grounded clues, and decides when to stop, forming a compact question-directed evidence sketch before answer generation. To optimize this behavior without human-annotated answers or rationale paths, MANGO samples multiple ANS rollouts and applies group-relative training with two rewards: answer preference from listwise self-ranking and path consistency from stable ordered panel trajectories. The combined reward is optimized with group-relative policy training, encouraging the model to improve both final answers and the panel-level evidence paths that support them. Experiments on standard manga understanding benchmarks show that MANGO achieves state-of-the-art performance across different settings.

1 Introduction

We study manga VQA, the task of answering natural-language questions about manga pages. Unlike conventional visual question answering over natural images, where the relevant evidence often lies within a single scene, manga pages present a structured visual narrative. Information is distributed across ordered panels, embedded text, character appearances, gestures, and implicit transitions between events (Baek et al., 2026). A correct answer may therefore depend on identifying relevant panels, linking dialogue to recurring characters, and reconstructing temporal or causal relations expressed through the page layout and narrative flow. This makes manga VQA less a problem of recognizing isolated visual cues than of understanding distributed, question-dependent evidence in a compact multimodal story.

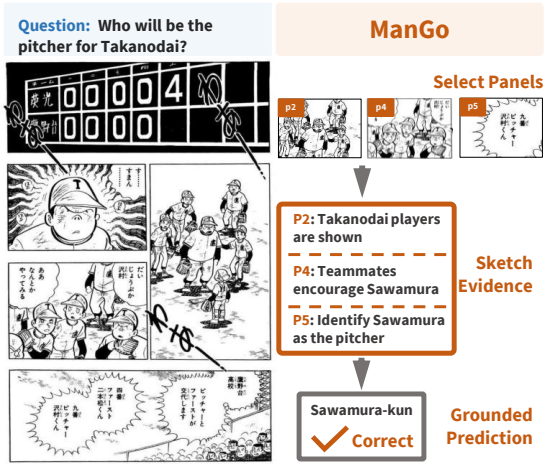

Figure 1: An illustration of MANGO. Rather than answering directly from the full manga page, MANGO actively selects question-relevant panels, writes compact panel-grounded clues, and composes them into a narrative evidence sketch. ©Mikio Yoshimori

This form of evidence distribution makes active evidence-seeking a natural fit for manga VQA. Active reasoning differs from passive problem solving in that the model must identify and acquire task-relevant evidence before solving the task (Zhou et al., 2025). For manga pages, this means deciding which panel to inspect, what clue to retain, and when the accumulated evidence is sufficient for answering. This framing avoids passively encoding the whole page once, and instead matches how manga VQA requires question-relevant evidence

*Equal Contribution.
†Corresponding Authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 15858–15873
October 24-29, 2026 ©2026 Association for Computational Linguistics

to be progressively discovered and connected.

Motivated by this view, we propose MANGO (Manga Active Narrative Grounding Optimization), an unsupervised framework for active manga VQA. MANGO consists of two coupled components: Active Narrative Sketching (ANS) for panel-level evidence construction, and group-relative training for unsupervised policy optimization. As shown in Fig. 1, given a manga page, panel boxes, and a question, ANS constructs a structured evidence sketch before answer generation. At each step, the model performs active panel perception by selecting a panel under the current question and sketch state, extracting a concise clue grounded in that panel, and deciding whether to continue or answer. The resulting narrative evidence sketch is ordered by the evolving evidence need rather than by page order alone, and may revisit a panel when it provides a different useful clue. In this way, ANS turns manga VQA into an iterative process of question-guided panel inspection and evidence-state update.

Prompting alone is unreliable for learning such active sketching behavior. However, direct supervision is difficult to scale because the model requires process-level supervision, not only final-answer labels. Such supervision must specify the sequence of evidence-seeking decisions along the reasoning path: which panel to inspect, which clue to preserve, how to avoid irrelevant details, and when the sketch is sufficient for answering. We therefore train MANGO with unsupervised reward modeling and group-relative policy optimization, using questions and panel annotations without relying on human-annotated answers or rationale paths.

To obtain training signals without gold-answer supervision, MANGO derives rewards from relations among sampled rollouts. For each question, the policy samples multiple ANS rollouts, each containing an evidence sketch and a final answer. Then, we construct two complementary rewards. First, answer preference reward estimates answer quality by merging duplicate final answers into candidate clusters and asking the policy model to rank them under the same manga page and question, producing a dense answer-level signal beyond frequency or surface-form agreement. Second, path consistency reward evaluates whether a rollout's ordered panel trajectory is supported by stable evidence paths from high-preference rollouts. The two rewards are combined and optimized with group-relative policy optimization, encouraging the policy to improve both the final answer and the panel-level evidence path that supports it.

We evaluate MANGO on standard manga understanding benchmarks and show that it achieves state-of-the-art performance across open-ended and multiple-choice settings. Overall, our work makes three contributions: (1) we propose MANGO, an unsupervised framework for active manga visual question answering; (2) we introduce ANS, which combines active panel perception with narrative evidence sketching to construct compact, question-guided panel evidence paths; and (3) we develop an unsupervised reward modeling scheme with answer preference and path consistency rewards, enabling group-relative training without human-annotated answers or rationale paths.

2 Background and Preliminaries

Active Perception. Active perception studies how an agent actively acquires task-relevant information instead of passively encoding a fixed observation. Classical active vision formulates perception as a closed-loop process, where the agent decides what, where, how, or when to sense according to its current goal and observations (Aloimonos et al., 1988). Recent multimodal reasoning methods instantiate this idea in MLLMs by allowing the model to select regions, crop images, or zoom into details before answering (Sarch et al., 2025; Zheng et al., 2026; Shen et al., 2025).

Given an input image I and question q, active perception can be abstracted as a sequence of task-conditioned perceptual actions:

$$a_t^p \sim \pi_\theta(\cdot \mid I, q, h_{t-1}), \qquad h_t = h_{t-1} \circ o_t, \quad (1)$$

where a_t^p is the perceptual action at step t, o_t is the acquired observation, and h_t is the accumulated perception history. In manga VQA, the perceptual action is naturally defined over panels. We therefore instantiate active perception as question-guided panel inspection, where each action selects one panel and updates a compact evidence sketch.

Group-Relative Policy Optimization. Reinforcement learning has become a common post-training strategy for improving LLM reasoning (Yu et al., 2025a; Gao et al., 2025; Zheng et al., 2025; Shrivastava et al., 2026). Group Relative Policy Optimization (GRPO) is a critic-free variant of PPO that estimates advantages by comparing multiple responses sampled for the same prompt, avoiding the need to train a separate value model (Shao et al., 2024). Given an input x, GRPO samples a group

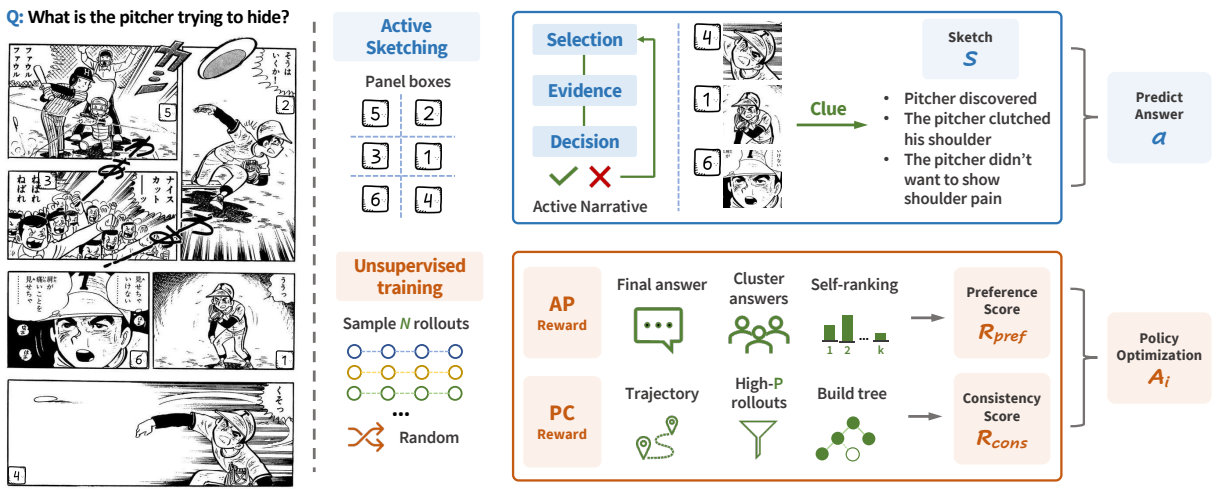

Figure 2: Overview of MANGO. The upper part shows Active Narrative Sketching, where the model performs panel selection, evidence sketching, and stopping decisions to build a compact sketch S before answering. The lower part shows unsupervised training: sampled rollouts are scored by answer preference and path consistency rewards, which are combined for group-relative policy optimization. ©Mikio Yoshimori

of responses $\{Y_i\}_{i=1}^{G}$ from the old policy and computes a normalized advantage from their rewards:

$$Y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \qquad A_i = \frac{R_i - \mu_R}{\sigma_R + \delta}, \quad (2)$$

where R_i is the reward of response Y_i, and μ_R, σ_R are the mean and standard deviation of rewards within the group.

The policy is then updated with a PPO-style clipped objective over generated tokens:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{i,t}\left[\min\left(\rho_{i,t}A_i, \bar{\rho}_{i,t}A_i\right)\right], \quad (3)$$

where $\rho_{i,t}$ is the token-level probability ratio between the current and old policies, and $\bar{\rho}_{i,t}$ is its clipped version. GRPO is well suited to our setting because manga VQA naturally permits multiple sampled reasoning trajectories for the same question. We therefore use group-relative optimization with rewards derived from sampled answers and their panel-level evidence paths, rather than from annotated answers or rationales.

3 MANGO

We propose MANGO, an unsupervised reasoning framework for manga visual question answering that learns to answer questions by actively sketching question-relevant evidence across panels, as shown in Fig. 2. It consists of Active Narrative Sketching for panel-level evidence construction and group-relative training for unsupervised policy optimization. Given a manga page I, panel boxes $B = \{b_m\}_{m=1}^{M}$, and a question q, the policy model π_θ generates a structured sketch and a final answer.

3.1 Active Narrative Sketching

Manga pages present visual and textual evidence as panel-based narratives, where answering a question often requires selectively following characters, dialogue, actions, and cross-panel relations. To model this process, we introduce Active Narrative Sketching (ANS), an active sketching procedure that decides *where to look*, *what to retain*, and *when to stop*, yielding a compact panel-grounded sketch before answer generation. ANS therefore connects perception and reasoning through a recurrent sketch state: each perception step adds one grounded clue, and the accumulated sketch serves as the evidence state for the final answer.

Active Panel Perception. Instead of passively answering from the whole page, ANS treats manga VQA as an iterative evidence-seeking process over panels. At each step, the policy conditions on the question, the manga page, and the current sketch state to perform three actions: (i) *panel selection*, which chooses a panel $p_t \in \{1, \ldots, M\}$ as the next visual narrative unit to inspect; (ii) *evidence sketching*, which extracts a concise clue c_t that is visually supported by the selected panel and relevant to the question; and (iii) *stopping decision*, which determines whether the accumulated sketch is sufficient for answering or another panel should be inspected. This active process adapts visual computation to the question, rather than treating all panels as an unordered set of image regions.

Formally, the sketch state is initialized as $S_0 = \emptyset$. At step t, the policy appends a panel-clue pair to

the sketch:

$$S_t = S_{t-1} \circ e_t, \qquad e_t = (p_t, c_t), \tag{4}$$

where p_t is the selected panel and c_t is the extracted clue. The process continues until the policy emits the stopping action, after which the accumulated sketch S_T is passed to the answer generation stage.

Narrative Evidence Sketch. The step-wise updates above form a structured evidence sketch

$$S = \big((p_1, c_1), (p_2, c_2), \ldots, (p_L, c_L)\big), \tag{5}$$

where p_t is the panel inspected at step t and c_t is a short clue grounded in that panel. The sketch follows a question-directed evidence order rather than the page reading order: the model may first inspect a panel containing key dialogue, then move to another panel to identify the speaker, and later revisit a previous panel to verify a character relation. Each clue is constrained to be *local*, *atomic*, and *question-relevant*: it should describe evidence visible in the selected panel, focus on one useful fact, and contribute to answering the current question.

To make this active sketching process explicit and parseable, we implement ANS with a structured generation interface, where each step outputs a selected panel, a grounded clue, and a continue-or-answer decision:

Active Sketching Prompt Template

Given a manga page with visible panel IDs and a question q, construct a compact evidence sketch before answering.

Step t:
Panel: select one panel ID p_t to inspect.
Clue: write one short clue c_t visible in the selected panel and relevant to q.
Decision: choose CONTINUE or ANSWER.

If ANSWER, generate the final answer from the collected clues.

Once the model decides to answer, the final response is generated as

$$\mathbf{a} \sim \pi_\theta(\mathbf{a} \mid I, B, q, S_T), \tag{6}$$

where $\mathbf{a}$ is the predicted answer sequence. For open-ended VQA, $\mathbf{a}$ is a natural-language response; for multiple-choice VQA, it is selected from the candidate option.

3.2 Unsupervised Reward Modeling

We optimize ANS without using human-annotated answers or rationale paths as rewards. For each manga page I, panel boxes B, and question q, we sample N rollouts from the current policy:

$$Y_i \sim \pi_\theta(\cdot \mid I, B, q), \qquad i = 1, \ldots, N, \tag{7}$$

where each rollout Y_i contains an active sketch and a final answer. From each rollout, we parse the final answer and the ordered panel trajectory

$$z_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,L_i}), \tag{8}$$

where L_i is the number of perception steps in Y_i. We then construct two complementary rewards: an answer preference reward that evaluates the sampled answer, and a path consistency reward that evaluates whether the panel trajectory is supported by stable evidence paths among sampled rollouts.

Answer Preference Reward. Since Manga VQA often admits open-ended answers, answer frequency or surface-form agreement may not reliably indicate answer quality. We therefore estimate answer preference through listwise self-ranking. For each valid rollout, we extract its final answer and merge duplicate answers into candidate clusters

$$\mathcal{A} = \{\alpha_1, \ldots, \alpha_K\}, \tag{9}$$

where each α_k denotes a unique answer and may be produced by multiple rollouts. Instances with $K < 2$ are skipped since no relative preference can be formed. We then ask the policy to rank the candidates under the same manga page and question, yielding a relative quality signal without annotated labels. With r_t^k denoting the rank position of α_k in the t-th sampled ranking, we compute

$$\begin{aligned} R_{\text{pref}}(\alpha_k) &= \frac{1}{T(K-1)} \sum_{t=1}^{T} (K - r_t^k), \\ R_{\text{pref}}(Y_i) &= R_{\text{pref}}(\alpha(Y_i)), \end{aligned} \tag{10}$$

where $\alpha(Y_i)$ is the answer cluster assigned to rollout Y_i. The normalized score lies in $[0, 1]$ and serves as the answer-level reward.

Path Consistency Reward. The answer reward scores the final response but does not directly evaluate whether the supporting panels form a stable reasoning path. We therefore introduce a path consistency reward over the parsed panel trajectories. Using the answer preference scores above, we select high-preference rollouts as reference paths and organize their trajectories into a prefix tree. Let

$\mathcal{H}$ denote rollouts with $R_{\text{pref}}(Y_i) \geq \tau$. For a prefix $u = (p_1, \ldots, p_\ell)$, its value is the average preference score of reference rollouts that contain u. The consistency reward of rollout Y_i is the average value of the prefixes along its own trajectory:

$$V(u) = \frac{\sum_{j \in \mathcal{H}} \mathbb{I}[u \preceq z_j] R_{\text{pref}}(Y_j)}{\sum_{j \in \mathcal{H}} \mathbb{I}[u \preceq z_j] + \epsilon},$$
$$R_{\text{cons}}(Y_i) = \frac{1}{L_i} \sum_{\ell=1}^{L_i} V(p_{i,1}, \ldots, p_{i,\ell}), \tag{11}$$

where $u \preceq z_j$ indicates that u is a prefix of trajectory z_j. Unlike set overlap, this reward is order-sensitive, allowing trajectories with the same panels but different orders to receive different scores. This is aligned with manga VQA, where answers often depend on ordered panels and cross-panel temporal or causal relations.

3.3 Group-Relative Training

During training, we randomly permute visible panel IDs while keeping panel boxes fixed to avoid ID-order shortcuts, and then optimize the policy using the rewards derived above.

Reward Composition. For each rollout, we combine answer quality and path consistency into a single training reward:

$$R(y_i) = R_{\text{fmt}}(y_i)\big(\lambda_{\text{pref}} R_{\text{pref}}(y_i) + \lambda_{\text{cons}} R_{\text{cons}}(y_i)\big), \tag{12}$$

where $R_{\text{fmt}}(y_i) \in \{0, 1\}$ checks whether the rollout follows the required ANS format, and λ_{pref} and λ_{cons} balance the answer-quality and path-consistency terms. The format reward prevents malformed outputs from receiving reward through accidental answer matches or panel overlaps. Thus, a rollout is favored only when it produces a plausible answer and supports it with a stable panel-level evidence path.

Policy Optimization. We optimize the policy with group-relative policy optimization. For each question, sampled rollouts form a group, and rollout rewards are normalized into advantages:

$$A_i = \frac{R(y_i) - \mu_R}{\sigma_R + \epsilon}. \tag{13}$$

Let $Y_i = (y_{i,1}, \ldots, y_{i,|Y_i|})$ denote the token sequence of rollout i. With the token-level ratio

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid y_{i,<t}, I, B, q)}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid y_{i,<t}, I, B, q)}, \tag{14}$$

we minimize

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_{i,t}\left[\min\left(\rho_{i,t} A_i, \bar{\rho}_{i,t} A_i\right)\right], \tag{15}$$

where $\bar{\rho}_{i,t} = \text{clip}(\rho_{i,t}, 1 - \epsilon, 1 + \epsilon)$. The self-ranking samples are used only for reward construction and are not directly optimized, avoiding reinforcement of the model's own judging behavior.

4 Experiments

4.1 Experimental Settings

Datasets and Benchmarks. For unsupervised training, we construct a question-only subset from the training set of MangaVQA (Baek et al., 2026). We first remove samples with empty questions/answers or identical question-answer texts and then randomly sample 512 questions, using only the questions and their corresponding Manga109 panel annotations (Aizawa et al., 2020); answer labels are not used for reward construction. For evaluation, we use three manga understanding benchmarks. (1) **MangaVQA** evaluates open-ended manga VQA with *Exact Extraction* and *Descriptive Answering* subsets, and we report both subset and overall scores. (2) **ChrOMIC** (Hou et al., 2026) evaluates chronological reasoning over comics; we use its Japanese manga subset and report both the original multiple-choice setting and a manually converted open-ended setting. For the latter, we retain only questions that admit free-form answers and use the text of the original correct option as the reference answer. (3) **MangaUB** (Ikuta et al., 2025) is a large-scale multiple-choice benchmark for manga understanding. We report results separately for single- and multi-panel questions, which respectively probe within-panel evidence extraction and cross-panel evidence aggregation, along with overall performance.

Evaluation Protocol. Following the official MangaVQA evaluation protocol, we employ Gemini 2.5 Flash (Comanici et al., 2025) (`gemini-2.5-flash`) as an LLM judge for open-ended evaluation. Given a question, a reference answer, and a model-generated response, the judge assesses the appropriateness and relevance of the response with respect to the reference and assigns a score from 1 to 10. We apply the same protocol to the open-ended ChrOMIC setting. For the multiple-choice ChrOMIC setting and the MangaUB evaluations, we report accuracy. Unless otherwise specified, all results are averaged over three sampled responses per question.

Models	MangaVQA			ChrOMIC		MangaUB		
	Extraction	Description	Overall	Multiple-choice	Open-ended	Single-panel	Multi-panel	Overall
General Perception								
MangaLMM‡	6.87	6.38	6.60	0.5108	3.83	0.6307	0.3958	0.5065
Molmo2-8B	2.71	2.86	2.79	0.4372	4.89	0.7547	0.5358	0.6389
InternVL3.5-8B	3.40	3.11	3.24	0.4978	5.40	0.7571	0.5092	0.6260
MiniCPM-V 4.5	4.41	4.40	4.24	0.5541	5.18	0.7647	0.4890	0.6189
Keye-VL-1.5-8B	3.57	2.65	3.07	0.4459	4.15	0.7624	0.4590	0.6020
LLaVA-OneVision-1.5	2.96	3.15	3.07	0.3030	5.34	0.7447	0.3741	0.5488
Kimi-VL-A3B-Instruct	4.32	3.51	3.88	0.3983	5.53	0.7530	0.4413	0.5882
Qwen2.5-VL-7B-Instruct	5.88	4.82	5.31	0.5298	5.54	0.7156	0.3826	0.5388
Qwen3-VL-8B-Instruct	7.18	5.74	6.40	0.4242	5.13	0.7568	0.4818	0.6114
Active Perception								
TreeVGR	4.41	3.09	3.69	0.4719	2.82	0.7005	0.4089	0.5464
DeepEyes	5.00	4.51	4.73	0.5195	5.67	0.7303	0.3984	0.5548
ViGoRL	4.33	3.72	3.99	0.4935	5.38	–	–	–
MANGO (Qwen2.5)	6.16	5.45	5.78	0.5368	5.60	0.7248	0.4197	0.5635
MANGO (Qwen2.5)‡	7.05	6.47	6.73	0.5671	6.06	0.7584	0.4535	0.5972
MANGO (Qwen3)	**7.20**	**6.64**	**6.90**	**0.5801**	**6.59**	**0.7711**	**0.5536**	**0.6561**

Table 1: Main results on three manga understanding benchmarks. MangaVQA is evaluated on *Exact Extraction*, *Descriptive Answering*, and overall performance; ChrOMIC is evaluated on *Multiple-choice* and *Open-ended* settings; and MangaUB is evaluated on *Single-panel*, *Multi-panel*, and overall performance. Models marked with ‡ are fine-tuned on MangaOCR data. Best results are in bold and second-best results are underlined.

Baselines. We compare MANGO with two baselines. (a) **General Perception** covers manga-specialized and general-purpose MLLMs (Baek et al., 2026; Clark et al., 2026; Wang et al., 2025a; Yu et al., 2025b; Yang et al., 2025a; An et al., 2025; Team et al., 2025; Bai et al., 2025b,a) that answer from a single-pass perception of the full page, without explicitly searching for intermediate visual evidence. (b) **Active Perception** includes methods that iteratively inspect visual regions or acquire visual evidence before prediction (Wang et al., 2026; Zheng et al., 2026; Sarch et al., 2025) which are based on Qwen2.5-VL.

Implementation Details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the base model and train it with MANGO on two NVIDIA A800 GPUs for 11 epochs, with a total training time of approximately 18 hours. We set the batch size to 8 and the learning rate to 4×10^{-6}. For each training instance, we sample 8 rollouts from the policy and query the policy 8 times for answer ranking. For reward modeling, the preference threshold is $\tau = 0.5$, and the reward weights are $\lambda_{\text{pref}} = 0.7$ and $\lambda_{\text{cons}} = 0.3$. Following recent observations that intrinsic unsupervised RL can become unstable when scaled aggressively (He et al., 2026), we use the compact training subset described above. During training, visible panel IDs are randomly permuted while panel boxes remain fixed to reduce ID-order shortcuts.

ANS	R_{pref}	R_{cons}	MangaVQA	ChrOMIC	
				MC	OE
✓	✗	✗	6.45	0.4285	5.18
✗	✓	✗	6.60	0.4372	6.12
✓	✓	✗	6.67	0.4545	6.31
✓	✓	✓	**6.90**	**0.5801**	**6.59**

Table 2: Ablation study on the main components of MANGO. R_{pref} denotes the answer preference reward, and R_{cons} denotes the path consistency reward.

4.2 Main Results

Tab. 1 compares MANGO with general perception and active perception baselines across three manga understanding benchmarks. Overall, MANGO achieves the strongest and most consistent performance across open-ended, multiple-choice, single-panel, and multi-panel settings. Compared with general perception models, MANGO benefits from explicitly decomposing manga VQA into panel-level evidence construction before answer generation, which is better aligned with the narrative structure of manga pages. Compared with active perception baselines, MANGO does not merely inspect visual regions; it retains selected evidence as a compact narrative sketch and updates the sketch through question-guided panel perception. These results indicate that MANGO is more effective on tasks that require linking characters, dialogue, and events across panels.

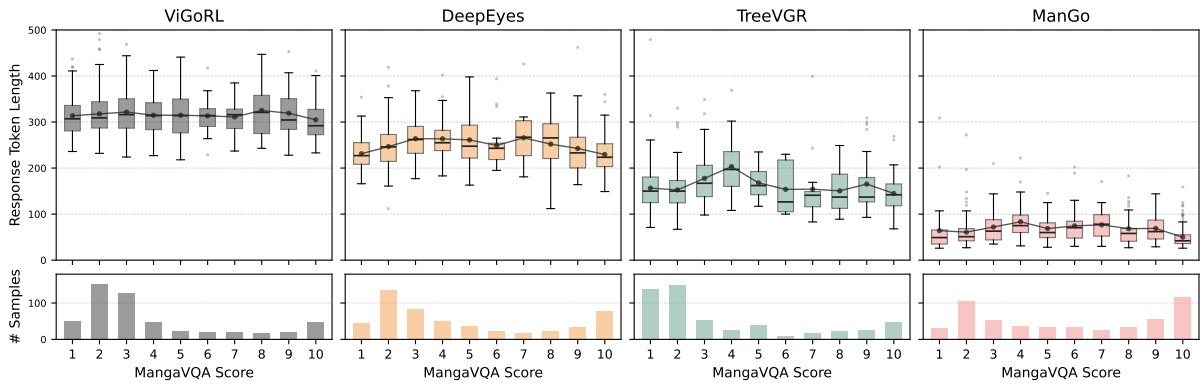

Figure 3: Response analysis on MangaVQA. Boxplots show the distribution of response token lengths within each score bin, and lower bars indicate the number of samples. Compared with other active perception methods, MANGO concentrates more samples in high scores while maintaining shorter responses.

Since MangaLMM incorporates MangaOCR supervision in addition to MangaVQA training, we also apply MANGO to a Qwen2.5-VL checkpoint that has been fine-tuned on MangaOCR under the same SFT protocol as MangaLMM. Combined with the improvements over the vanilla Qwen2.5-VL and Qwen3-VL baselines reported earlier, these results demonstrate that MANGO remains effective across different backbones and initializations.

4.3 Ablation Study

Component Ablation. Tab. 2 validates the contribution of each component. ANS alone provides a structured panel-grounded reasoning format, but without reward optimization its performance remains limited. R_{pref} improves answer quality, especially for open-ended evaluation, while combining it with ANS yields more stable gains by optimizing answers within a structured sketch. Adding R_{cons} further improves all settings, showing that path consistency complements answer preference by encouraging high-quality answers to follow stable panel trajectories. The full model performs best, confirming that active sketching, answer preference, and path consistency are beneficial.

Preference Reward Ablation. Tab. 3 compares answer preference estimators with ANS, path consistency, and policy optimization fixed. Clustering (Zhang et al., 2025) or confidence (Prabhudesai et al., 2025) based rewards provide weak supervision for manga VQA, where answer quality depends on fine-grained speaker, character, and event relations. POLAR (Dou et al., 2025) remains competitive but underperforms our self-ranking reward despite using ground-truth answers, indicating that reference matching can miss narrative distinctions. Self-ranking is more effective because it evaluates candidate answers relative to other rollouts under the same page and question, which better matches the group-relative training objective.

Preference Reward	GT	MangaVQA	ChrOMIC	
			MC	OE
POLAR	✓	6.35	0.5411	5.91
EMPO	✗	4.75	0.4978	2.67
RENT	✗	1.01	0.2554	1.00
Self-Ranking (ours)	✗	**6.90**	**0.5801**	**6.59**

Table 3: Ablation study on preference reward estimators, with ANS, path consistency, and policy optimization fixed. *GT* indicates whether ground-truth answers are used in reward computation.

Consistency Reward Ablation. Tab. 4 compares different estimators for the path consistency reward while keeping ANS, answer preference reward, and policy optimization fixed. Jaccard matching performs bad because it ignores panel order and only measures unordered overlap. Sequence-based metrics improve by considering ordered panel paths, showing that the structure of the evidence trajectory matters for manga. Prefix Tree performs best across all settings, suggesting that rewarding stable prefixes better captures how early panel choices guide subsequent evidence construction. This supports our design of path consistency reward, which stabilizes question-directed panel trajectories rather than merely matching selected panels.

4.4 Analysis

Evaluation Robustness. To examine whether the open-ended evaluation is sensitive to the choice of

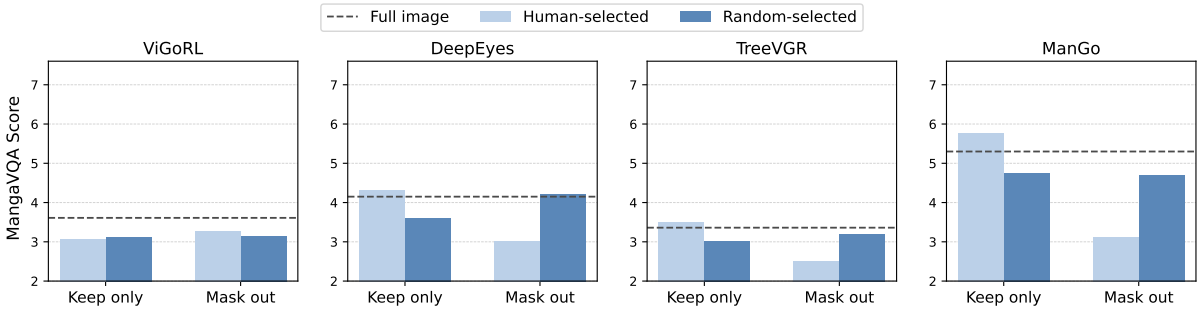

Figure 4: Panel perturbation analysis on MangaVQA. Each subplot compares manually annotated answer-supporting panels with size-matched random panels under keep-only and mask-out settings.

Consistency Reward	MangaVQA	ChrOMIC	
		MC	OE
Jaccard	6.70	0.4762	6.33
Longest Common Subsequence	6.76	0.5108	6.41
Longest Common Substring	6.79	0.5325	6.47
Prefix Tree (ours)	**6.90**	**0.5801**	**6.59**

Table 4: Ablation study on consistency reward estimators. All variants use the same ANS generation, answer preference reward, and policy optimization, differing only in how path consistency is computed.

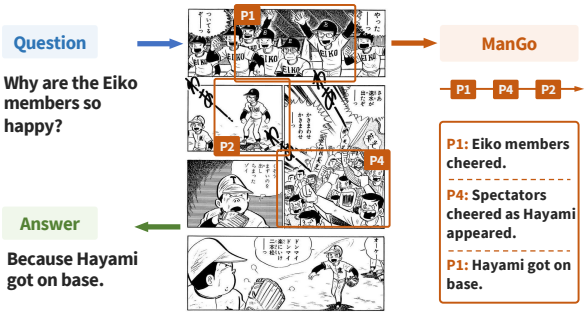

Figure 5: Case study of MANGO. Given a question about why the Eiko members are happy, MANGO selects supporting panels, sketches concise clues about the team's reaction and Hayami's action, and answers by linking the emotion to its narrative cause. ©Mikio Yoshimori

LLM judge, we rescored the same MangaVQA responses using GPT-4o (Hurst et al., 2024) as an alternative to Gemini 2.5 Flash. As shown in Tab. 5, although GPT-4o generally assigns lower absolute scores, the overall relative performance remains consistent, with MANGO achieving the highest score under both judges. These results support the reliability of the evaluation across judge models.

Response Compactness. Fig. 3 shows that existing active perception methods tend to produce substantially longer responses, suggesting that iterative visual exploration can introduce redundant

Models	Judge: Gemini	Judge: GPT-4o
MangaLMM	6.68	6.57
Molmo2-8B	2.79	2.15
InternVL3.5-8B	3.24	3.02
MiniCPM-V 4.5	4.24	3.98
Keye-VL-1.5-8B	3.07	2.38
LLaVA-OneVision-1.5	3.07	2.44
Kimi-VL-A3B-Instruct	3.88	3.52
Qwen2.5-VL-7B-Instruct	5.31	4.92
Qwen3-VL-8B-Instruct	6.40	6.03
TreeVGR	3.69	3.58
DeepEyes	4.73	4.55
ViGoRL	3.99	3.81
MANGO	**6.90**	**6.78**

Table 5: Robustness of LLM-as-a-judge evaluation on MangaVQA. The same model responses are scored by Gemini 2.5 Flash and GPT-4o respectively.

or weakly organized information. In contrast, MANGO maintains a compact response distribution while achieving stronger MangaVQA performance. This supports the design of Active Narrative Sketching: instead of accumulating lengthy visual descriptions, the model records local, atomic, and question-relevant clues in a structured sketch. The result suggests that manga VQA benefits more from concise evidence organization than from expanding the amount of intermediate text.

Evidence Utilization. Fig. 4 evaluates whether models rely on panels that actually support the answer, using keep-only and mask-out perturbations. Answer-supporting panels are manually annotated under a necessity criterion: a panel is labeled as supporting if its removal precludes derivation of the correct answer. Size-matched randomly selected panels are used as controls. MANGO benefits more when human-selected panels are retained and drops more when they are masked out, while random-selected panels lead to weaker or less consistent effects. This indicates that MANGO is more tightly

grounded in answer-relevant panels, rather than relying on incidental context or generic visual exploration. This result is consistent with our reward design: path consistency encourages high-quality answers to be supported by stable ordered panel trajectories, making the learned policy more sensitive to answer-supporting panels.

Case Study. Fig. 5 shows a representative reasoning path produced by MANGO. Given the question about why the Eiko members are happy, the model selects panels showing the cheering team, Hayami's appearance, and Hayami getting on base, then composes these clues into a compact sketch. The final answer follows from this panel-level narrative link rather than from a single salient visual cue, illustrating how MANGO grounds manga reasoning in ordered supporting evidence.

5 Related Work

Visual Question Answering. VQA has evolved from object- and attribute-centric question answering on natural images toward more complex multimodal understanding scenarios. Recent studies increasingly examine models in settings that involve text-rich scenes, multilingual visual content, scientific figures and ambiguous visual contexts (Tang et al., 2025; Jiu et al., 2025; Burgess et al., 2025; Jian et al., 2025; Yang et al., 2025b; Liu et al., 2021, 2024a,b; Xiao et al., 2025). This trend suggests that VQA is shifting from simple visual recognition to reasoning over structured visual information and task-specific contexts, of which manga VQA is a representative but less explored case.

Manga VQA. Existing manga and comic benchmarks have studied related abilities such as reading order, panel sequencing, speaker identification, and multi-panel understanding (Vivoli et al., 2024; Ikuta et al., 2025; Vivoli et al., 2025; Wang et al., 2025b; Hou et al., 2026). MangaVQA and MangaLMM introduce a manga-specific VQA benchmark and specialized model for manga understanding (Baek et al., 2026). Nevertheless, compared with general VQA and other specialized VQA domains, manga VQA remains relatively underexplored, especially for open-ended question answering over panel-structured narratives.

6 Conclusion

We present MANGO, an unsupervised framework for manga visual question answering that treats answering as active panel-level evidence seeking. Through Active Narrative Sketching, the model selects relevant panels, extracts concise grounded clues, and builds a compact evidence sketch before answering. With answer preference and path consistency rewards optimized by group-relative training, MANGO learns this behavior without human-annotated answers or rationale paths. Experiments on multiple manga understanding benchmarks show that MANGO improves both final answer accuracy and evidence grounding.

Limitations

Although MANGO achieves strong performance on existing manga understanding benchmarks, its evaluation is still constrained by the limited availability of large-scale manga VQA resources. Current benchmarks cover only a portion of the broad spectrum of manga comprehension, with data sources that remain limited in genre, era, language, and cultural context. Therefore, the reported results may not fully characterize model behavior across more diverse manga scenarios. Future work could develop broader benchmarks and evaluation protocols to support a more comprehensive assessment of manga understanding systems.

Ethics Statement

This work studies manga VQA using existing manga understanding benchmarks for academic research. Manga pages are copyrighted creative works, and we follow the usage conditions of the corresponding datasets. We do not redistribute raw manga images beyond what is permitted by the original resources. All manual annotations used in this work were conducted by the authors and disagreements are resolved through majority voting. No external annotators, crowdworkers, or human-subject participants were recruited, and no annotator personal data were collected. Since manga may contain culturally specific expressions, sensitive themes, or subjective narrative interpretations, model predictions should not be treated as authoritative readings of the original works. We encourage future use of this work to respect dataset licenses, avoid unauthorized redistribution of copyrighted materials, and conduct additional human verification when applying manga VQA systems in user-facing scenarios.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62332010 and by the Fundamental Research Funds for the Central Universities under Grant No. CUC26TD06.

References

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

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. 1988. Active vision. *International journal of computer vision*, 1(4):333–356.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. 2026. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5349–5370.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19552–19564.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28652–28668.

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

Shihan Dou, Shichun Liu, Yuming Yang, Yicheng Zou, Yunhua Zhou, Shuhao Xing, Chenhao Huang, Qiming Ge, haijun Lv, Demin Song, Songyang Gao, Chengqi Lyu, Enyu Zhou, Honglin Guo, Zhiheng Xi, Qipeng Guo, Wenwei Zhang, Tao Gui, Qi Zhang, and 3 others. 2025. Pre-trained policy discriminators are general reward models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 170177–170217. Curran Associates, Inc.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. 2025. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, et al. 2026. How far can unsupervised rlvr scale llm training? *arXiv preprint arXiv:2603.08660*.

Bingxuan Hou, Jiayi Lin, Chenyang Zhang, Dapeng Yin, Shuyue Zhu, Qingqing Hong, Mengna Gao, and Junli Wang. 2026. Chromic: Chronological reasoning across multi-panel comics. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4384–4400.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. 2025. Mangaub: A manga understanding benchmark for large multimodal models. *IEEE MultiMedia*, 32(2):33–43.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638.

Sha Jiu, Yu Weng, Mengxiao Zhu, Chong Feng, Zheng Liu, and Jialedongzhu. 2025. TVQACML: Benchmarking text-centric visual question answering in

multilingual Chinese minority languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13957–13967, Suzhou, China. Association for Computational Linguistics.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134.

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024a. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5753–5762.

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2024b. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *Transactions on Machine Learning Research*.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael Tarr, Aviral Kumar, and Katerina Fragkiadaki. 2025. Grounded reinforcement learning for visual reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 150977–151013. Curran Associates, Inc.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6602–6618, Suzhou, China. Association for Computational Linguistics.

Vaishnavi Shrivastava, Ahmed H Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. 2026. Sample more to think less: Group filtered policy optimization for concise reasoning. In *International Conference on Learning Representations*, volume 2026, pages 98439–98462.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, et al. 2025. Mtvqa: Benchmarking multilingual text-centric visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7748–7763.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. 2024. Comix: A comprehensive benchmark for multi-task comic understanding. *Advances in Neural Information Processing Systems*, 37:140828–140846.

Emanuele Vivoli, Artemis Llabrés, Mohamed Ali Souibgui, Marco Bertini, Ernest Valveny Llobet, and Dimosthenis Karatzas. 2025. Comicspap: understanding comic strips by picking the correct panel. In *International Conference on Document Analysis and Recognition*, pages 337–350. Springer.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. 2026. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, volume 2026, pages 148769–148794.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Qingxiu Dong, Rui Li, Yixin Yang, Yifan Pu, Weiyao Luo, Yiru Wang, Xiangdi Meng, Wenjie Li, and Zhifang Sui. 2025b. Beyond single frames: Can LMMs comprehend implicit narratives in comic strip? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6436–6452, Suzhou, China. Association for Computational Linguistics.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. 2025. Egoblind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. 2025a. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*.

Shuo Yang, Caren Han, Siwen Luo, and Eduard Hovy. 2025b. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16967–16986.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025a. DAPO: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 113222–113244. Curran Associates, Inc.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. 2025b. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 67345–67372. Curran Associates, Inc.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. 2026. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 126775–126798.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. 2025. From passive to active reasoning: Can large language models ask the right questions under incomplete information? In *International Conference on Machine Learning*, pages 78714–78758. PMLR.

A Prompting Framework

MANGO uses a two-stage prompting framework during policy optimization. The first stage samples ANS rollouts, where the policy produces a compact panel-grounded evidence sketch followed by a final answer. The second stage estimates answer-level preference by asking the policy model to rank distinct candidate answers sampled from the first stage. Gold answers are not provided in either stage.

A.1 Active Narrative Sketching Prompt

Fig. 6 shows the prompt used to sample ANS rollouts from the policy model. Given a manga page with visible panel IDs and a question, the model is instructed to actively trace evidence across panels before answering. Each reasoning step associates one panel ID with one concise clue, making the generated sketch directly parseable into panel-clue pairs. The prompt also discourages page-level summaries and unrelated visual descriptions, so that the intermediate reasoning remains compact and question-directed.

The output is parsed into two fields. The content inside `<think>` is used to extract the ordered panel trajectory, while the content inside `<answer>` is treated as the final answer and later used for self-ranking.

A.2 Rollout Answer Ranking Prompt

Fig. 7 shows the prompt used to estimate answer-level preference without gold answers. For each training instance, we first extract the final answers from sampled ANS rollouts and merge duplicate answers into distinct candidate options. The policy is then asked to rank these options from best to worst according to answer relevance, visual support, completeness, and conciseness. The resulting rankings are aggregated with Borda count to produce the answer preference reward.

This stage produces a relative ordering over candidate answers rather than an absolute correctness label. We sample multiple rankings to reduce ranking noise and obtain dense preference scores for all candidate answers.

B Benchmark & Models

B.1 Manga Understanding Benchmark

We evaluate model performance on three manga understanding benchmarks: MangaVQA, ChrOMIC, and MangaUB. The detailed information about these benchmarks is as follows:

- MangaVQA (Baek et al., 2026): A benchmark designed to evaluate multimodal large language models on contextual visual question answering over manga pages. Built on images from Manga109 (Aizawa et al., 2020), it consists of 526 manually curated and verified question–answer pairs that require models to jointly understand visual content, panel layout, character interactions, and embedded textual information. The benchmark covers both single-panel and multi-panel reasoning, and includes questions requiring either exact text extraction or descriptive contextual interpretation.

Active Narrative Sketching Prompt

You are answering a manga VQA question by actively tracing evidence across labeled panels.
Goal:
Read the page like a manga reader who investigates the answer step by step. Instead of selecting all useful panels first, build an evidence chain as the reasoning unfolds.
Each step links one panel id to one concise clue, and each new clue should fill a missing piece of evidence.
How to trace evidence:
1. Start with the panel most directly related to the question.
2. From the current panel, extract one concise clue that helps answer the question.
3. Decide what evidence is still missing, such as speaker, action, dialogue meaning, emotion, object, location, time, cause, or result.
4. Move to the same panel or another panel only if it provides a different clue needed to fill that gap.
5. Stop once the collected clues are sufficient to answer the question. Important:
- A panel may be used multiple times if it provides different necessary clues.
- Do not describe unrelated details or summarize the whole page.
- Use only visible evidence from the manga page.
- Keep the final answer brief and do not include the evidence chain in it.
Question:
{question}
Output format:
<think>
pX: clue → pY: clue → . . .
</think>
<answer>brief final answer</answer>

Figure 6: Prompt used for Active Narrative Sketching.

- ChrOMIC (Hou et al., 2026): A benchmark designed to evaluate chronological reasoning in multi-panel comics. Built from public comic datasets, it contains 998 comic pages spanning Western and Japanese styles. The benchmark defines three complementary tasks including panel reordering, description Reordering and multiple-choice question answering. Through a human–AI collaborative annotation pipeline with manual refinement and filtering, ChrOMIC provides a focused testbed for assessing whether LVLMs can move beyond static image understanding toward layout-aware, temporally grounded narrative reasoning in comics.

- MangaUB (Ikuta et al., 2025): A manga understanding benchmark built on Manga109 to evaluate large multimodal models across single- and multi-panel scenarios. It contains 6,585 questions and 18,179 prompts covering both low-level scene recognition and higher-level narrative understanding, including tasks such as location, time, weather, character counting, emotion recognition, panel localization, and next-panel prediction. By combining expert annotations with multiple-choice prompt variants, MangaUB provides a compact testbed for assessing models' visual, contextual, and sequential understanding of manga.

B.2 General Perception Models

We evaluated and measured both manga-specialized and general-purpose multimodal large language models (MLLM) on the above benchmarks:

- **MangaLMM** (Baek et al., 2026): A manga-specialized VLM introduced together with MangaVQA. It is adapted from Qwen2.5-VL and fine-tuned on MangaOCR annotations and synthetic manga VQA data, making it the most closely related domain-specific baseline for manga visual question answering.

Rollout Answer Ranking Prompt

You are judging candidate answers for a manga VQA question.
Rank the options from best to worst. A better answer should:
1. Directly answer the question.
2. Be fully supported by the manga image.
3. Include all necessary information without extra explanation.
4. Avoid unsupported details, vague wording, and reasoning text.
Important:
- Do not reward an answer for being longer.
- If two answers are similarly correct, prefer the more concise and direct one.
- Penalize extra details that are not needed or not clearly supported by the image.
Question: {question}
Options:
{option_lines}
Output:
Return all option letters exactly once, from best to worst, separated by commas.

Figure 7: Prompt used for Rollout Answer Ranking.

- **Molmo2-8B** (Clark et al., 2026): An open-weight VLM from Ai2 designed for image, multi-image, and video understanding with grounding capabilities. It uses an efficient packing and message-tree encoding recipe, bidirectional attention over vision tokens, and a token-weighting strategy to improve multimodal grounding and video-language understanding.
- **InternVL3.5-8B** (Wang et al., 2025a): An open-source MLLM from the InternVL3.5 family. Its training pipeline combines multimodal continual pre-training, supervised fine-tuning, and Cascade Reinforcement Learning, with the goal of improving multimodal reasoning, document/OCR understanding, multi-image comprehension, and general visual-language instruction following.
- **MiniCPM-V-4.5** (Yu et al., 2025b): An efficient general-purpose MLLM from the MiniCPM-V series. It is built on Qwen3-8B and SigLIP2-400M, and introduces a unified 3D-Resampler for compact image and video encoding, together with training strategies for document understanding, OCR, and both short- and long-form reasoning.
- **Keye-VL-1.5-8B** (Yang et al., 2025a): A general-purpose multimodal model developed by the Kwai Keye Team. It uses Qwen3-8B as the language decoder and a SigLIP-initialized vision encoder, and supports native dynamic resolution, SlowFast video encoding, and 3D RoPE for unified image, video, and text processing.
- **LLaVA-OneVision-1.5** (An et al., 2025): A fully open-source VLM from the LLaVA-OneVision-1.5 family. It is trained with a released end-to-end framework and large-scale curated pretraining and instruction-tuning data, supporting native-resolution image understanding and general multimodal instruction following.
- **Kimi-VL-A3B-Instruct** (Team et al., 2025): An efficient open-source MoE VLM developed by Moonshot AI. It activates only a small subset of language-model parameters during inference while supporting multimodal reasoning, long-context understanding, and agent-oriented capabilities.
- **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025b): A general-purpose VLM from the Qwen2.5-VL series. It improves visual recognition, object localization, document parsing, long-video understanding, and visual-agent capabilities, and serves as the backbone for several active visual reasoning baselines in our experiments.
- **Qwen3-VL-8B-Instruct** (Bai et al., 2025a): A general-purpose VLM from the Qwen3-

VL series. It supports long interleaved multimodal contexts and introduces architectural updates such as Interleaved-MRoPE, DeepStack, and text-timestamp alignment to improve spatial-temporal modeling, fine-grained vision-language alignment, and video temporal grounding.

B.3 Active Perception Models

We further evaluated models that go beyond general perception by explicitly selecting or inspecting visual regions during reasoning:

- **TreeVGR** (Wang et al., 2026) is initialized from Qwen2.5-VL-7B and trained with a two-stage pipeline consisting of cold-start supervised fine-tuning and reinforcement learning. It introduces traceable visual evidence by supervising both answer correctness and bounding-box localization quality, encouraging the model to ground its reasoning in relevant image regions.
- **DeepEyes** (Zheng et al., 2026) is built on Qwen2.5-VL-7B and learns active visual reasoning through end-to-end reinforcement learning. During reasoning, it can ground and crop relevant image regions, forming an interleaved multimodal reasoning process that supports fine-grained visual inspection.
- **ViGoRL** (Sarch et al., 2025) is also based on Qwen2.5-VL and trains visually grounded reasoning with reinforcement learning. It anchors reasoning steps to explicit image coordinates and uses a multi-turn visual feedback mechanism to dynamically zoom into predicted regions for fine-grained perception.

B.4 Software and Inference Settings

For MANGO, we implement training with PyTorch 2.8.0, Transformers 4.57.0, Verl 0.7.1, and vLLM 0.11.0. For baseline models, we follow their official repositories and use their recommended software environments, preprocessing pipelines, and default inference configurations unless otherwise specified.

C Extended Results

C.1 Inference Efficiency

To further quantify inference efficiency, we measure per-example inference latency on MangaVQA using a single NVIDIA A800 GPU with a batch size of 1. As shown in Tab. 6, MANGO requires 5.21 seconds per example, lower than all existing active-perception baselines. In particular, it is substantially faster than DeepEyes and ViGoRL, which incur additional overhead from iterative cropping and visual feedback. Although MANGO performs iterative panel selection and evidence construction, its latency remains comparable to that of single-pass models while delivering stronger MangaVQA performance. This favorable efficiency–performance trade-off is consistent with the compact design of ANS, which retains only concise, question-relevant clues during inference.

Models	Latency (s)
General Perception	
Qwen3-VL-8B-Instruct	8.09
LLaVA-OneVision-1.5	3.46
Kimi-VL-A3B-Instruct	4.16
Active Perception	
TreeVGR	5.88
DeepEyes	30.02
ViGoRL	70.49
MANGO	**5.21**

Table 6: Per-example inference latency on MangaVQA, measured on a single NVIDIA A800 GPU with a batch size of 1. The best result among active-perception methods is shown in bold.

C.2 Comparison with Closed-Source Models

To provide broader reference points, we additionally evaluate several strong closed-source models using the same evaluation protocol. As shown in Tab. 7, MANGO remains competitive despite using an 8B open-source backbone, outperforming several closed-source models on both benchmarks. While the strongest closed-source models achieve higher overall performance, these results demonstrate that MANGO provides a competitive alternative through targeted post-training for manga narrative understanding.

C.3 Evidence Behavior of ANS

The process analysis in the main text compares MANGO with other active perception methods. Here, we further focus on the behavior of ANS itself and examine whether its generated sketches capture useful evidence and organize it in an effective way.

Panel-level evidence relevance. We first test whether the panels selected by ANS are actually

Models	MangaVQA	ChrOMIC (MC)
Claude Sonnet 4.5	5.84	0.4762
Claude Sonnet 5	6.48	0.6017
GPT-5.2	6.59	0.5281
GPT-5.4 mini	6.25	0.4459
GPT-5.4	7.98	0.5714
Gemini 2.5 Flash	7.26	0.5974
Gemini 3.5 Flash	**8.33**	**0.7229**
MANGO	6.90	0.5801

Table 7: Comparison with strong closed-source models under the same evaluation protocol. MangaVQA open-ended responses are scored from 1 to 10 using the LLM judge, while ChrOMIC reports accuracy in its original multiple-choice setting.

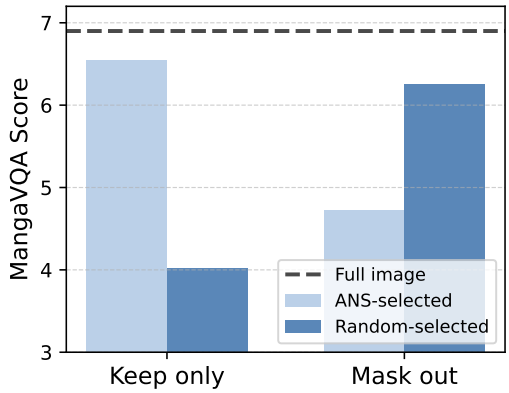

Figure 8: Counterfactual panel perturbation analysis on MangaVQA. We compare ANS-selected and randomly selected panels under keep-only and mask-out settings. The dashed line denotes the full-image score.

useful for answering. As shown in Fig. 8, keeping only ANS-selected panels yields much better performance than keeping the same number of randomly selected panels. Conversely, masking ANS-selected panels leads to a larger degradation than masking random panels. These two perturbations show that ANS-selected panels are strongly associated with answer-relevant evidence. Notably, the generated evidence path contains only 2.4 panels on average out of 11.3 panels per page, indicating that ANS preserves substantial answer-related information with a compact panel subset.

Evidence sketch accumulation. Beyond selecting relevant panels, we examine whether the organization of ANS affects answer generation. For model outputs on MangaVQA whose sketch contains four steps, we keep the steps in their original order or reorder them into reverse and random orders. For each order, we truncate the sketch after the first k steps ($k = 1, \ldots, 4$), append the resulting prefix to the original prompt, and ask the model to generate the final answer.

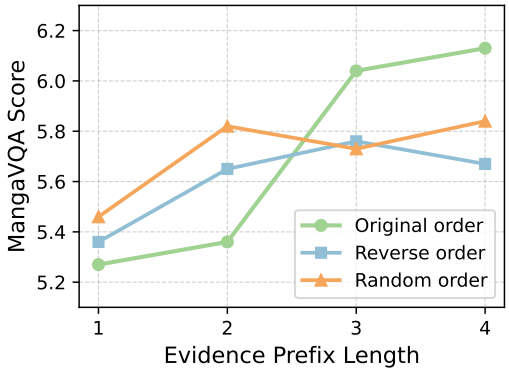

Figure 9: Evidence sketch accumulation analysis. We reorder length-4 evidence sketch generated by ANS into original, reverse, and random orders, progressively reveal their prefixes, and report the MangaVQA score at each prefix length.

As shown in Fig. 9, the original ANS produces a steadily improving curve as more steps are revealed, while the reverse and random variants saturate or fluctuate. Since all variants contain the same panel-clue steps, the difference cannot be explained by evidence selection alone. Instead, it suggests that ANS is useful not only because it collects relevant clues, but also because it arranges them into a question-directed evidence flow that better supports manga narrative understanding.