CHROMIC: Chronological Reasoning Across Multi-Panel Comics

Bingxuan Hou [1,2], Jiayi Lin [1,2], Chenyang Zhang [1,2], Dapeng Yin [1,2], Shuyue Zhu [1,2], Qingqing Hong [1,2], Mengna Gao [1,2], Junli Wang [1,2]*

[1] Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai 201804, China.

[2] National (Province-Ministry Joint) Collaborative Innovation Center for Financial Network Security, Tongji University, Shanghai 201804, China.

{2432023, 2331908, inkzhangcy, 2432122}@tongji.edu.cn

{2432272, 2332012, 2432121, junliwang}@tongji.edu.cn

Abstract

Large-scale vision–language models (LVLMs) have achieved remarkable progress on various reasoning tasks. However, most studies focus on natural photographic images and pay limited attention to multi-panel visual narratives such as comics. This leaves a clear gap in our understanding of how well LVLMs perform chronological reasoning across comic panels. To address this, we introduce CHROMIC, a new benchmark dataset for **chro**nological reasoning in multi-panel **comic**s. It covers six types of reasoning questions and spans both Western and Japanese comic styles. To ensure high-quality annotations, we customized a human–AI collaborative annotation process tailored to the characteristics of the two comic styles. We further introduce three core tasks: *Description Reordering* and *Panel Reordering*, which jointly assess models' ability to understand chronological order in panel sequences, and *Multiple-Choice Question Answering (MCQA)*, which evaluates narrative-level reasoning. We evaluate a range of open-source and commercial LVLMs on CHROMIC, and find that even the leading models struggle with panel-based chronological reasoning. Further analysis reveals key limitations, including weak visual action understanding and frequent hallucinations in fine-grained visual interpretation. [1]

1 Introduction

In recent years, the reasoning capabilities of LVLMs have attracted significant attention from researchers. Advanced LVLMs such as GPT-4o (OpenAI et al., 2024b) and Qwen (Qwen et al., 2025) have demonstrated remarkable performance

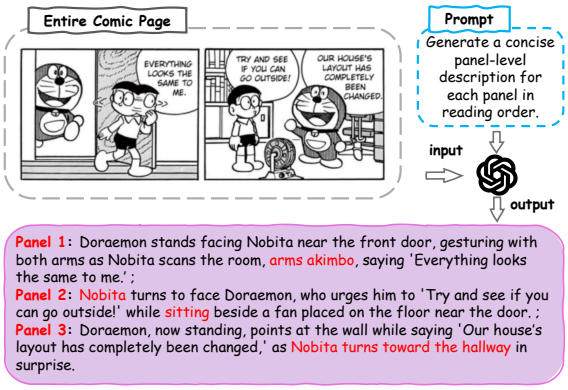


Figure 1: An error case produced by GPT-4o when directly generating panel-level descriptions for an entire comic page without any preprocessing. The model shows inaccuracies in understanding the comic page, including misjudging panel order, misidentifying details, and miscounting the number of panels. Errors are highlighted in red text.

improvements across a wide range of reasoning tasks. Among these, multi-image reasoning (Liu et al., 2024; Yeh et al., 2022; Huang et al., 2024; Jiang et al., 2024a) has become a prominent research focus, and recent efforts have increasingly turned toward the more complex domain of multi-panel visual understanding (Fan et al., 2024; Wang et al., 2024a).

Although some studies have begun to explore the capacity of LVLMs to understand and reason about multi-panel visual content in real-world scenarios, current research still faces notable limitations. Specifically, we identify two critical research gaps: (i) Lack of abstract images in evaluation: Current cross-modal training corpora predominantly focus on realistic photographs (Radford et al., 2021; Zhang et al., 2024a), lacking sufficient examples of abstract or exaggerated artistic styles (Zhang et al., 2024b; Li et al., 2024), resulting in very limited benchmarks for evaluating abstract-style images. (ii) Limited layout and chronological reasoning in

*Corresponding author. This work was supported by the National Key Research and Development Program of China under Grant 2022YFB4501704.

[1] Our benchmark is publicly available at https://github.com/daydayup586/ChrOMIC

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4384–4400
March 24-29, 2026 ©2026 Association for Computational Linguistics

multi-panel pages: Most LVLMs utilize autoregressive architectures, which restrict their capacity for layout comprehension. As shown in recent studies (Sachdeva and Zisserman, 2025; Meng et al., 2024), given that these models primarily rely on forward-sequential attention mechanisms, they perform well on clearly sequentialized image data but struggle to effectively interpret spatial and chronological relationships among multiple panels within a single page.

To address these limitations, we focus our research on an artistic form—comic (Iyyer et al., 2017). In comic understanding tasks, models must deduce the chronological sequence of multiple panels arranged spatially on a single page and further interpret the underlying narrative. Although some existing studies (Wu et al., 2024; Tang et al., 2025) have facilitated the understanding of comics, this reasoning paradigm remains largely underexplored, and current LVLMs still fall short, as illustrated in Figure 1.

To fill this gap, we propose three novel tasks: (i) **Panel Reordering**, which tests whether models can recover the correct sequential flow from shuffled panel images; (ii) **Description Reordering**, which evaluates a model's ability to reconstruct chronological order from textual panel descriptions; and (iii) **Multiple-Choice Question Answering**, which probes narrative-level inference across panels. Building on these tasks, we present CHROMIC, the first high-quality benchmark explicitly targeting chronological reasoning in multi-panel comics. The dataset comprises 998 comic pages with 6,700 panels and a total of 1,900 annotated instances, covering both Western and Japanese styles across diverse reading orders.

We conduct extensive experiments on CHROMIC with a wide range of commercial and open-source LVLMs. The results demonstrate that CHROMIC presents major challenges for LVLMs, highlighting persistent weaknesses in visual action understanding and narrative reasoning, pointing to important directions for future research.

2 Related Work

2.1 Multimodal Benchmarks

Large-scale vision–language models have attracted significant attention, and many benchmarks have been proposed to evaluate emerging tasks. For example, Das et al. (2024) introduces a multimodal multilingual benchmark, while Chen et al. (2024) and Wang et al. (2024b) focus on multimodal hallucination detection. Other studies address the scarcity of abstract image data with benchmarks on geometric shapes and scientific charts (Zhang et al., 2024b; Li et al., 2024), though comic benchmarks exist, deeper investigations into comic understanding remain limited. Multi-image benchmarks such as Liu et al. (2024); Fan et al. (2024); Huang et al. (2024); Jiang et al. (2024b); Song et al. (2025) mainly emphasize sequential image understanding, often overlooking specialized multi-panel formats like comics. Although MultipanelVQA (Fan et al., 2024) is the first benchmark targeting multi-panel inputs, it largely consists of images without explicit chronological relations (e.g., web pages (He et al., 2024)) and focuses on static layouts rather than panel-level chronological reasoning.

In contrast, we propose CHROMIC, a benchmark centered on panel-level chronological reasoning in comics, filling this gap and offering a new perspective on multimodal sequential reasoning beyond static image comprehension.

2.2 Comic Understanding Research

Research on comic understanding has progressed from early visual tasks to more complex reasoning. Prior work has explored dialogue prediction, next-panel forecasting, and related datasets (Iyyer et al., 2017; Agrawal et al., 2023). Transcription-oriented studies further decompose the problem into subtasks such as panel segmentation, dialogue segmentation, and OCR (Vivoli et al., 2024; Sachdeva and Zisserman, 2024; Sachdeva et al., 2024; Rigaud et al., 2024), later extended to dialogue behavior recognition (Martinek et al., 2024).

Beyond transcription, researchers have investigated description generation to support storyline comprehension (Ramaprasad, 2023; Wang et al., 2024a), comic completion to infer missing content (Guo et al., 2023), and LVLM-based comic generation with resources such as MangaZero (Wu et al., 2025). Benchmarks like MangaUB (Ikuta et al., 2024) target single-panel understanding, while YESBUT (Hu et al., 2024) examines non-linear reasoning without explicit chronology.

In contrast, our work presents the first VQA benchmark for multi-panel comics, explicitly targeting chronological reasoning and narrative understanding, filling an important research gap.

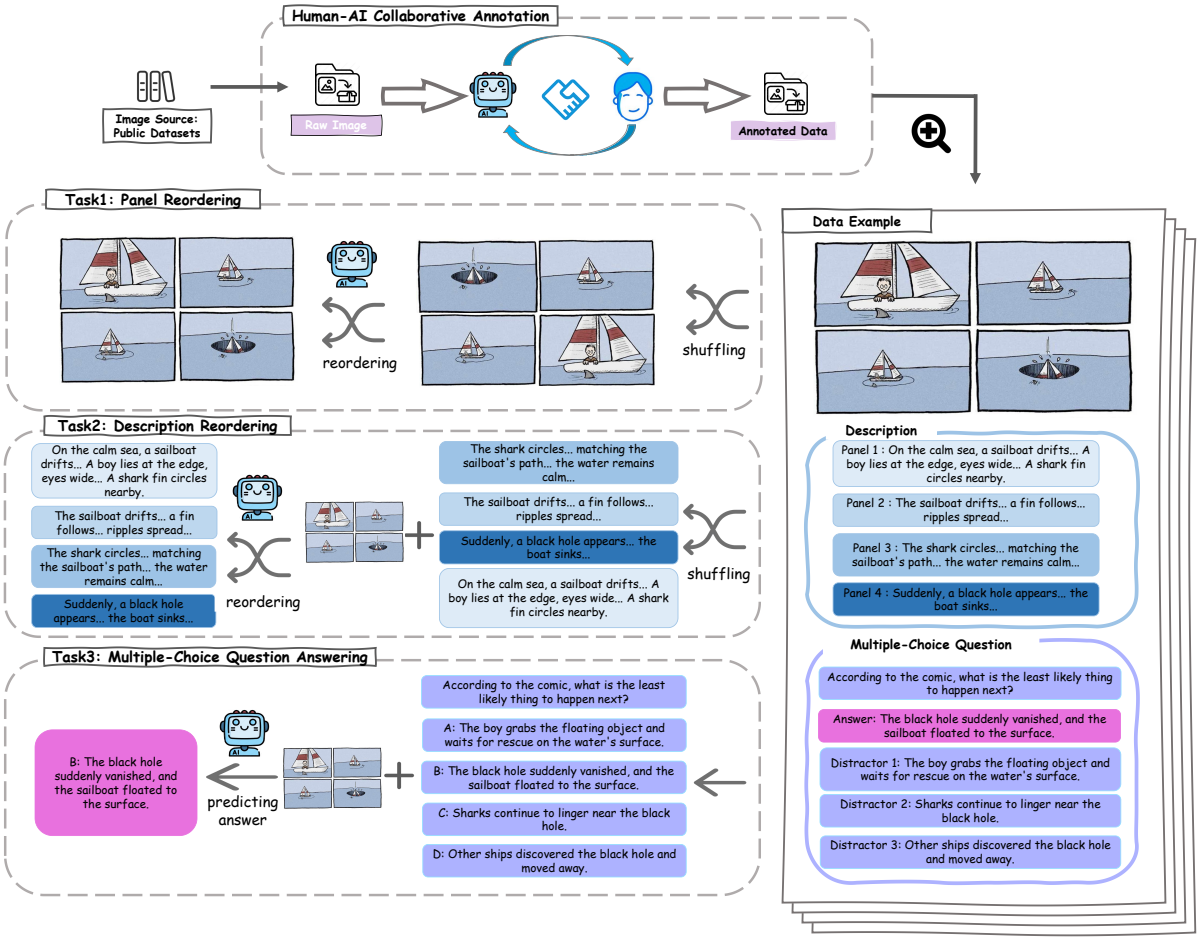


Figure 2: **Overview of the CHROMIC benchmark.** CHROMIC is constructed via a human–AI collaborative pipeline using images from public comic datasets. Each instance provides a comic page with associated annotations (panel descriptions and/or a reasoning question). We design three tasks: *Panel Reordering* and *Description Reordering*, both targeting panel-level chronology, and *Multiple-Choice Question Answering*, which evaluates narrative-level inference across panels.

3 CHROMIC

As illustrated in Figure 2, we design and construct CHROMIC, with formal task definitions provided in Section 3.1 and dataset construction details described in Sections 3.2 and 3.3.

3.1 Task Design

To evaluate the chronological reasoning capabilities of LVLMs in comic narratives, we introduce three tasks: (*i*) *Panel Reordering*, (*ii*) *Description Reordering* and (*iii*) *Multiple-Choice Question Answering*. These tasks respectively assess the model's ability to *recognize panel order* and *perform deeper narrative reasoning*. A comparison with prior comic benchmarks is provided in Table 1, highlighting that CHROMIC uniquely combines chronological reasoning with multimodal narrative understanding.

Dataset	Chron.	Reason.	Multi.
Manga109 (Fujimoto et al., 2016)	✗	✗	✗
COMSET (Agrawal et al., 2023)	✗	✓	✓
COMICS (Iyyer et al., 2017)	✗	✓	✓
YESBUT (Hu et al., 2024)	✗	✓	✓
CHROMIC (ours)	✓	✓	✓

Table 1: Comparison of comic-related datasets. **Chron.** = requires chronological understanding; **Reason.** = requires reasoning; **Multi.** = requires multimodal input.

3.1.1 Panel Reordering

As shown in the "Task 1" region of Figure 2, let $\mathcal{I}$ be a comic page with an ordered panel sequence $\boldsymbol{P}^\star = (p_1^\star, \ldots, p_n^\star)$, where n represents the number of panels. We sample a random permutation $\pi \in S_n$ and form the shuffled sequence $\tilde{\boldsymbol{P}} = \left(p_{\pi(1)}^\star, \ldots, p_{\pi(n)}^\star\right)$, which is concatenated into a composite image $\tilde{\mathcal{I}}$. Given $\tilde{\mathcal{I}}$, the model

predicts a permutation $\hat{\pi} \in S_n$ and is correct when:

$$\tilde{\boldsymbol{P}}_{\hat{\pi}} = \boldsymbol{P}^{\star}. \tag{1}$$

3.1.2 Description Reordering

As shown in the "Task 2" region of Figure 2, reusing the panel sequence $\boldsymbol{P}^{\star}$ from above, each panel $p_i^{\star}$ is paired with a textual description $d_i^{\star}$, yielding $\boldsymbol{d}^{\star} = (d_1^{\star}, \ldots, d_n^{\star})$. We then apply a permutation $\pi \in S_n$ to obtain a randomly shuffled list $\tilde{\boldsymbol{d}} = \left(d_{\pi(1)}^{\star}, \ldots, d_{\pi(n)}^{\star}\right)$ and provide $(\mathcal{I}, \tilde{\boldsymbol{d}})$ as input. The objective is to predict $\pi_i \in S_n$ such that the re-ordered list $\tilde{\boldsymbol{d}}_{\pi_i}$ completely matches the ground-truth $\boldsymbol{d}^{\star}$, namely:

$$\tilde{\boldsymbol{d}}_{\pi_i} = \boldsymbol{d}^{\star}. \tag{2}$$

3.1.3 Multiple-Choice Question Answering

As illustrated in the "Task 3" region of Figure 2, let $(\mathcal{I}, q, C)$ be the input, where q denotes the question and $C = \{c_1, c_2, c_3, c_4\}$ is the set of candidate answers and $a^{\star} \in \{1, 2, 3, 4\}$ is the ground-truth index. The model predicts $\hat{a} \in \{1, 2, 3, 4\}$. With direct matching evaluation, the single-example accuracy is:

$$\mathcal{A}_{\text{single}} = \mathbb{1}\left[\hat{a} = a^{\star}\right], \tag{3}$$

where the indicator $\mathbb{1}[\hat{a} = a^{\star}]$ equals 1 if the predicted answer $\hat{a}$ matches the ground-truth answer $a^{\star}$, and 0 otherwise. For a set of M samples we report the mean accuracy:

$$\mathcal{A}_{\text{mean}} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\left[\hat{a}_i = a_i^{\star}\right]. \tag{4}$$

3.2 Image Selection

The comic images in CHROMIC are sourced from three publicly available datasets, covering diverse reading orders as shown in Table 2. We carefully curate all images to ensure quality, diversity, and clear panel segmentation, prioritizing comics with well-defined reading orders and filtering out inappropriate or harmful content. To avoid bias from low-resource languages, we include only English-language comics, aligning with the training data of most current LVLMs.

Specifically, we use images from the Mementos dataset (Wang et al., 2024a) to construct the dialogue-free Western subset, from the ComSet dataset (Agrawal et al., 2023) to build the dialogue-rich Western subset, and from the MangaZero dataset (Wu et al., 2025) to extract double-page spreads for the Japanese subset. The panel count distribution is shown in Figure 5 in Appendix C.

Dataset	Top-to-Bottom	Left-to-Right	Right-to-Left
COMICS	✗	✓	✗
Manga109	✗	✗	✓
Mementos (Wang et al., 2024a)	✓	✓	✗
COMSET	✗	✓	✗
MangaZero (Wu et al., 2025)	✓	✗	✓
CHROMIC (ours)	✓	✓	✓

Table 2: Comparison of supported reading orders across comic datasets.

3.3 Human-AI Collaborative Annotation

Inspired by Hu et al. (2024), we adopt a human–AI workflow to support our three tasks. For each comic image, an LVLM first drafts task materials (e.g., shuffled panels/descriptions, MCQ candidates); trained annotators then refine these outputs to produce the gold standard. Not every example supplies all three task views, as filtering is applied for quality and relevance.

Six professionally trained annotators participate in the process, representing diverse gender and age groups, all with at least a bachelor's degree. The end-to-end pipeline is illustrated in Figure 3.

3.3.1 Textual Description Annotation

Image Preprocessing To ensure accurate panel segmentation and ordering, we design a dedicated pipeline. Specifically, we train a YOLOv12-based detector (Tian et al., 2025) on a curated set of annotated panels to handle diverse comic styles. Each page layout is represented as a Directed Acyclic Graph (DAG), and panel sequences are determined via topological sorting with distinct rules for Western (left-to-right) and Japanese (right-to-left) conventions, producing an ordered panel list $\boldsymbol{P}^{\star} = (p_1^{\star}, \ldots, p_n^{\star})$. Based on these sequences, we apply style-specific preprocessing: in Western comics, visual prompts are inserted to explicitly mark panel positions and reading order, while in Japanese comics, panels are cropped and reordered into individual images to reduce layout ambiguity.

Description Annotation We adopt a human–AI collaborative strategy to generate panel-level descriptions $\boldsymbol{d}^{\star}$. GPT-4o (OpenAI et al., 2024b) is used for Western comics and Qwen-Max (Qwen et al., 2025) for Japanese comics. Generation is guided by the preprocessed panel inputs and specialized prompts that enforce panel count consistency and chronological flow. Human annotators then refine the drafts, correcting errors, enhanc-

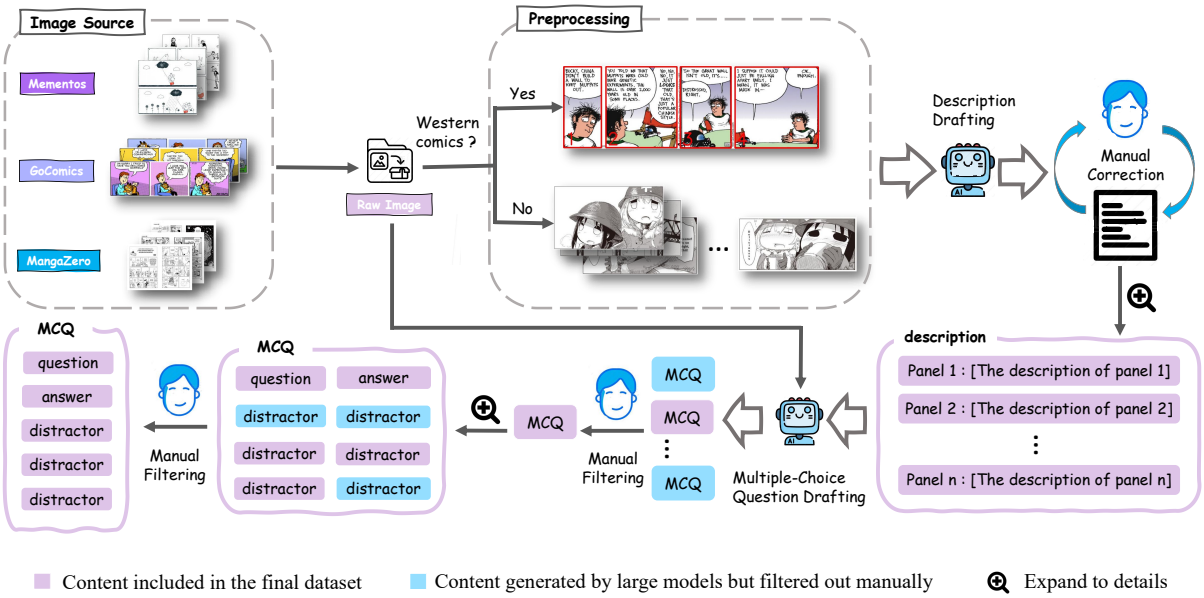


Figure 3: Dataset construction pipeline. In the preprocessing stage: for Western comics, visual prompts are added; for Japanese comics, the pages are segmented into image sequences. After manually selecting the best Multiple-Choice Question (MCQ), the model generates three additional distractors. Annotators then select the three most misleading ones from a pool of six candidates to finalize the MCQ.

ing clarity, and ensuring alignment with visuals. This hybrid strategy ensures both scalability and annotation reliability.

3.3.2 Data Construction for Panel Reordering

We build the *Panel Reordering* set with an image-only pipeline: panels are detected and cropped, then randomly shuffled and concatenated into a composite test image. To simplify evaluation and reduce ambiguity, we restrict sources to pages with at most four panels and include a brief manual audit to verify the ground-truth panel order before shuffling. No textual annotation or post-editing is required; each instance is released with the shuffled composite and the target permutation.

	WC	JC	WCND	Overall
Number of pages	446	234	118	795
Avg. panels	4.4	10.9	5.1	6.4
Avg. question tokens	14.1	15.6	16.3	14.8
Avg. option tokens	10.9	12.2	12.6	11.5

Table 3: Task 3 (MCQA) data statistics for three subsets: **WC** (Western comics), **JC** (Japanese comics, right-to-left), and **WCND** (Western comics without dialogue).

3.3.3 Multiple-Choice Question Annotation

We design a three-stage strategy that guides LVLMs to generate multiple-choice questions for assessing chronological understanding in comics, combined with human refinement.

In the first stage, the annotation model generates a question q, an answer $a^\star$, and three distractors $C_{\text{init}} = \{c_1^{\text{init}}, c_2^{\text{init}}, c_3^{\text{init}}\}$ based on six specified perspectives: *(1) Cross-panel action tracing (CPT)*, *(2) Temporal character action sequences (CAS)*, *(3) Inferred motivations between non-consecutive panels (IMN)*, *(4) Implicit character motives (ICM)*, *(5) Prediction beyond shown panels (PBP)*, and *(6) Counterfactual reasoning (CFR)*. A detailed description of these perspectives is provided in Appendix A.4. Human annotators then review and select high-quality, narrative-relevant questions.

In the second stage, given the filtered question–answer pair, the model produces three additional highly deceptive distractors $C_{\text{adv}} = \{c_1^{\text{adv}}, c_2^{\text{adv}}, c_3^{\text{adv}}\}$ with explanations. Annotators then choose the three most misleading distractors from $C_{\text{init}} \cup C_{\text{adv}}$ and combine them with the correct answer $a^\star$ to construct the final candidate set $C = \{a^\star, c_1, c_2, c_3\}$.

In the third stage, to ensure that vision grounding is indispensable, we conduct multiple rounds of text-only filtering to remove questions that could be answered without images, retaining only those that require genuine multimodal reasoning. This yields an evaluation that more faithfully reflects image–text integration. Dataset statistics for this third task (MCQA) are summarized in Table 3.

Model	Size	Acc	MAE	SRC	NDCG@n
Commercial LVLMs					
GPT-4o	–	**0.3735**	1.8396	0.2628	0.8044
Moonshot-V1	–	0.3621	1.8371	**0.2732**	0.7977
Qwen-Max	–	0.3480	1.9033	0.2157	0.8072
Gemini-2.0	–	0.3359	1.9344	0.1998	0.8175
Grok-3	–	0.3335	1.9032	0.2324	0.7859
Open-source LVLMs					
ahm-Phi-3.5-VI	4B	0.2641	2.0543	0.1445	**0.8963**
Gemma-3	4B	0.2924	1.9281	0.2125	0.7851
	12B	0.3175	1.8893	0.2463	0.7927
	27B	0.3288	1.9128	0.2218	0.8100
LLaMA-3.2-VI	11B	0.2402	**1.4037**	0.0320	0.8951
	90B	0.3180	1.9335	0.1959	0.8164
Qwen-2.5-VL	3B	0.2851	1.9531	0.2329	0.7944
	7B	0.3219	1.9414	0.2093	0.7792
	32B	0.3649	1.8545	0.2506	0.8070

Table 4: Performance comparison of models on the description reordering task. **Size** denotes parameter scale (e.g., 7B = 7 billion), and **"VI"** indicates vision-language instruction tuning. Here, n is the number of panels in each comic instance. Bold values indicate the best-performing model in each category.

Model	Size	Acc	MAE	SRC	NDCG@n
Commercial LVLMs					
GPT-4o	–	**0.3991**	**0.9724**	**0.1895**	0.6719
Moonshot-V1	–	0.2713	1.1103	0.0970	**0.7244**
Qwen-Max	–	0.2894	1.1165	0.0767	0.6154
Gemini-2.0	–	0.3484	1.0301	0.1519	0.6442
Grok-3	–	0.2901	1.1140	0.0789	0.6219

Table 5: Performance on the Panel Reordering (image-only) task.

4 Experiment

4.1 Experimental Setup

We conduct a systematic evaluation of several mainstream LVLMs, covering both the latest commercial closed-source models and open-source models. The commercial models evaluated include GPT-4o (OpenAI et al., 2024a), Moonshot-V1 (Team et al., 2025c), Qwen-Max (Bai et al., 2025), Gemini-2.0 (Team et al., 2025a), Grok-3 (xAI, 2025). The open-source models include ahm-Phi-3.5-VisionInstruct (Abdin et al., 2024), LLaMA-3.2-VisionInstruct (Meta AI, 2024), Gemma-3 (Team et al., 2025b) and Qwen-2.5-VL (Qwen et al., 2025).

To fairly assess each model's performance on panel sequence reasoning tasks, we standardize the input format for all evaluations: each model receives the full comic page image along with a text-formatted question. See Section 3.1 and Appendix F for details. All models are tested using their respective default settings.

Model	Size	WC	JC	WCND	Overall
Commercial LVLMs					
GPT-4o	–	**0.517**	0.463	**0.527**	**0.507**
Moonshot-V1	–	0.314	0.342	0.446	0.396
Qwen-Max	–	0.339	0.346	0.439	0.397
Gemini-2.0	–	**0.517**	**0.506**	0.460	0.482
Grok-3	–	0.441	0.420	0.464	0.448
Open-source LVLMs					
ahm-Phi-3.5-VI	4B	0.364	0.377	0.379	0.376
Gemma-3	4B	0.322	0.320	0.377	0.352
	12B	0.373	0.329	0.455	0.406
	27B	0.441	0.381	0.413	0.408
LLaMA-3.2-VI	11B	0.347	0.329	0.386	0.364
	90B	0.407	0.368	0.482	0.438
Qwen-2.5-VL	3B	0.424	0.294	0.357	0.348
	7B	0.356	0.351	0.368	0.361
	32B	0.432	0.316	0.404	0.382

Table 6: Model performance on different types of comics for the multiple-choice question answering task. The **Overall** column presents the average performance across all comic types.

4.2 Result Evaluation

Description Reordering Task For the Description Reordering task, we evaluate models using four metrics: **Accuracy** (whether each panel is placed in the correct position), **MAE** (the average deviation between predicted and true positions), **Spearman Rank Correlation** (overall ranking consistency), and **NDCG@n** (ranking quality of the top n key panels). The aggregated results appear in Table 4, where commercial models such as GPT-4o and Qwen-Max consistently outperform open-source alternatives. In this experiment, n is set to the actual number of panels on each comic page to comprehensively assess ordering performance. For detailed definitions and explanations of these evaluation metrics, please refer to Appendix B.1.

GPT-4o achieves the best overall performance. Among open-source models, Qwen-2.5-VL-32B performs competitively with commercial ones, while smaller models such as ahm-Phi-3.5-VisionInstruct and LLaMA-3.2-11B-VisionInstruct excel on NDCG@n, indicating that lightweight models can still be effective when ranking the most critical panels.

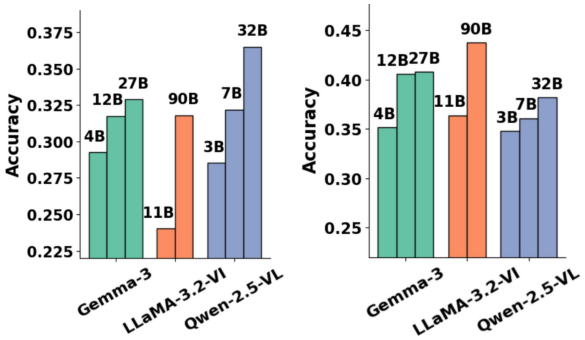
Panel Reordering Task On the image-only panel reordering task, the results in Table 5 show that GPT-4o achieves the highest accuracy, closely followed by Gemini-2.0. Both outperform others significantly, showing stronger ability to reconstruct narrative flow directly from visual evidence. However, most models still struggle with the task, highlighting the challenges of purely visual sequen-

Model	Size	CPT	CAS	IMN	ICM	PBP	CFR
Commercial LVLMs							
GPT-4o	–	**0.527**	**0.542**	0.569	0.492	0.439	**0.476**
Moonshot-V1	–	0.382	0.413	0.376	0.409	0.382	0.400
Qwen-Max	–	0.400	0.432	0.349	0.383	0.349	0.429
Gemini-2.0	–	0.509	0.465	**0.578**	**0.503**	0.447	0.381
Grok-3	–	0.500	0.510	0.459	0.461	0.366	0.362
Open-source LVLMs							
ahm-Phi-3.5-VI	4B	0.391	0.374	0.376	0.425	0.293	0.371
Gemma	4B	0.364	0.406	0.339	0.332	0.350	0.314
	12B	0.373	0.419	0.367	0.446	0.374	0.429
	27B	0.436	0.387	0.413	0.430	0.398	0.371
LLaMA-3.2-VI	11B	0.373	0.432	0.394	0.389	0.325	0.219
	90B	0.473	0.484	0.440	0.352	**0.463**	0.457
Qwen-2.5-VL	3B	0.327	0.335	0.413	0.342	0.293	0.400
	7B	0.318	0.361	0.358	0.399	0.350	0.352
	32B	0.409	0.432	0.330	0.383	0.350	0.371

Table 7: Performance across six chronological reasoning subcategories in multi-panel comics. **CPT**: Cross-panel tracing, **CAS**: Character action sequences (chronological reconstruction), **IMN**: Inferred motivations (non-consecutive panels), **ICM**: Implicit character motives, **PBP**: Prediction beyond panels (what happens next?), **CFR**: Counterfactual reasoning (what if X panel changed?).

Method	Acc of Task 1	Acc of Task 3
Human evaluation	0.9532	0.679
GPT-4o	**0.3735**	**0.507**
Gemini-2.0	0.3359	0.482

Table 8: Results of human evaluation compared with top-performing LVLMs. Task 1 = Description Reordering, Task 3 = Multiple-Choice QA.



(a) (b)

Figure 4: Accuracy comparison of open-source models at varying scales in Description Reordering task (a) and Multiple-Choice Question Answering task (b)

tial reasoning and layout comprehension.

Multiple-Choice Question Answering Task For the Multiple-Choice Question Answering task, we report accuracy, with results summarized in Tables 6 and 7. Overall, commercial LVLMs such as GPT-4o and Gemini-2.0 clearly outperform open-source counterparts, confirming the advantage of large-scale closed-source systems. However, performance on Japanese comics is generally weaker, particularly for smaller models, reflecting the challenges posed by denser layouts and longer panel sequences.

Table 7 shows a clear advantage for commercial LVLMs. GPT-4o delivers the strongest overall results—leading in **CPT** and **CAS** and achieving the best **CFR**—indicating robust panel tracing, action sequencing, and counterfactual reasoning. Gemini-2.0 complements this by topping **IMN** and **ICM**, suggesting superior implicit-motivation and causal inference. By contrast, open-source systems generally lag behind; the main exception is LLaMA-3.2-90B-VI, which attains the best **PBP** score, while Qwen-2.5-VL-32B is competitive on **CAS/ICM**. Overall, the results highlight both the dominance of commercial models and the subtask-specific complementarities that different architectures exhibit.

Human Evaluation Study To further validate the benchmark, we conducte a human evaluation by uniformly sampling 165 examples across different domains and difficulty levels. Four annotators independently complete both the *Description Reordering* and *Multiple-Choice QA* tasks. Results are summarized in Table 8.

Human annotators consistently outperform the best-performing LVLMs on both tasks. The gap is especially large in *Description Reordering*, where humans achieve near-perfect accuracy. This indicates that chronological reasoning and holistic narrative understanding remain challenging for current

Model	Def.	Txt-only	Mask Dial.
GPT-4o	0.3735	0.3203	0.3632
Moonshot-V1	0.3429	0.3180	0.2947
Qwen-Max	0.3480	0.3450	0.3324
Qwen-2.5-VL(7B)	0.3219	0.2708	0.2908
Claude-3.7-Sonnet	0.3821	0.3597	0.3911
DeepSeek-R1	—	0.3373	—

Table 9: Ablation results for *Description Reordering*. **Def.** = Default multimodal input; **Txt-only** = images removed; **Mask Dial.** = dialogues in images and dialogue/connective words in text descriptions masked.

Style	Direct	Auto (Preproc.)	Human-Refined
WC	1.3	3.5	4.7
JC	1.6	4.3	4.4

Table 10: Human evaluation of description faithfulness across comic styles. **Direct**: model-only generation; **Auto (Preproc.)**: automatically generated with preprocessing; **Human-Refined**: human post-edited after preprocessing.

LVLMs, underscoring the value of our benchmark for advancing multimodal sequential reasoning.

4.3 Analysis and Discussion

4.3.1 Scale and Design Effects

Experiments show that performance within the same model family generally improves with increasing parameter size, suggesting that scaling enhances chronological and narrative reasoning. However, models with comparable sizes can still exhibit notable performance gaps, likely due to differences in architectural design and training strategies, such as visual–textual pretraining balance and task-specific instruction tuning.

As illustrated in Figure 4a and Figure 4b, these trends highlight that both scale and architecture jointly shape model effectiveness, and that increasing size alone is not sufficient without careful design choices.

4.3.2 Multimodal Necessity Ablations

To ensure that our tasks truly require visual grounding, we conducte ablation experiments summarized in Table 9. For *Description Reordering (DR)*, two variants are tested: a pure text-only setting and a masked-dialogue setting where dialogue and connective words are removed while in-image dialogues are also masked. Both led to performance drops, with the text-only condition showing the sharpest decline, indicating that models rely more on reasoning over the full visual narrative than on matching dialogues or connective cues. Meanwhile, the decrease is not as significant as one might expect—a trend explainable by the task's inherent difficulty. As shown in Table 8, even with multimodal input, model performance remains far below human accuracy (0.9532), leaving little room for further degradation.

For *Multiple-Choice Question Answering*, we filtered out questions answerable by text alone, retaining only those that require genuine multimodal reasoning. The results confirm that visual input is indispensable for both tasks.

4.3.3 Annotation Reliability

To validate dataset reliability, we conducte an additional human evaluation. Two independent annotators, not involved in the initial labeling, assesse the alignment between textual descriptions and panels across dimensions such as character actions, scene elements, dialogue themes, and chronological indicators (see Appendix D). We evaluate three description types: (i) model-generated, (ii) automatically generated with preprocessing, and (iii) human-refined after preprocessing.

Results show that while automatic generation provides a useful starting point, human refinement is essential for improving fidelity and consistency. Table 10 presents quantitative results, and semantic similarity analysis further supports the necessity of human intervention. These findings demonstrate the robustness of our annotation pipeline and highlight the value of combining automation with human refinement.

5 Conclusion

In this work, we presente CHROMIC, a high-quality benchmark for structured visual narrative understanding in comics. The benchmark evaluates both panel-level chronology and narrative inference through three complementary tasks: *Panel Reordering*, *Description Reordering*, and *Multiple-Choice Question Answering*.

Our experiments show that CHROMIC presents significant challenges for current state-of-the-art LVLMs, and analysis reveals persistent limitations in visual action reasoning and story-level comprehension. These results highlight the need for models that are more layout-aware and sensitive to narrative structure. We expect CHROMIC to serve as a valuable resource for advancing research on multimodal narrative reasoning.

Limitations

The current scale of CHROMIC remains relatively modest compared to modern large-scale vision-language benchmarks. Although we intentionally included both Western and Japanese comic styles, other popular formats, such as vertically scrolling webtoons or Chinese comics, have not been covered. This omission could potentially limit the generalizability of our findings across different comic styles. Additionally, despite employing a human-AI collaborative annotation pipeline along with multi-stage quality assurance processes, residual annotation noise may persist, given that human annotators' subjective interpretations of comic content can vary.

From an ethical and social perspective, while we have manually filtered overtly offensive or harmful content, subtle cultural stereotypes or sensitive themes may still remain undetected. Furthermore, our current reliance on English descriptions and questions may inadvertently marginalize non-English comic communities.

In future work, we plan to expand the scale and stylistic diversity of CHROMIC and release multilingual annotations. Additionally, we will introduce further redundancy checks or expert verification processes to continually enhance annotation quality and reliability.

Ethics Statement

Data Copyright and Licensing. All data samples used in this work are collected from either publicly released datasets or publicly accessible content on social media platforms. To ensure compliance with copyright regulations, we do not redistribute raw comic images. Instead, we provide annotations and reference links to the original sources, allowing readers to access the materials without violating copyright protections. Before publishing any samples, we manually review the content to identify and remove potentially offensive, harmful, or inappropriate material.

Human Annotation. Our annotation process involved six human annotators who were responsible for reviewing and refining model-generated outputs. We ensured fair compensation by paying annotators an average hourly wage of $9 USD, aligning with common ethical standards for crowd or contract work. All annotators participated voluntarily and were informed of the nature and goals of the task.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Harsh Agrawal, Aditya Mishra, Manish Gupta, and Mausam. 2023. Multimodal persona based generation of comic dialogs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14164, Toronto, Canada. Association for Computational Linguistics.

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. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. 2024. Muffin or Chihuahua? challenging multimodal large language models with multipanel VQA. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6845–6863, Bangkok, Thailand. Association for Computational Linguistics.

Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Manga109 dataset and creation of metadata. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*. ACM.

Hongcheng Guo, Boyang Wang, Jiaqi Bai, Jiaheng Liu, Jian Yang, and Zhoujun Li. 2023. M2C: Towards

automatic multimodal manga complement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9876–9882, Singapore. Association for Computational Linguistics.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. WebVoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6864–6890, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024. Cracking the code of juxtaposition: Can ai models understand the humorous contradictions. *Preprint*, arXiv:2405.19088.

Shizhou Huang, Bo Xu, Changqun Li, Jiabo Ye, and Xin Lin. 2024. MNER-MI: A multi-image dataset for multimodal named entity recognition in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11452–11462, Torino, Italia. ELRA and ICCL.

Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. 2024. Mangaub: A manga understanding benchmark for large multimodal models. *Preprint*, arXiv:2407.19034.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, III, and Larry S. Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chaoya Jiang, Hongrui Jia, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2024a. Maven: An effective multi-granularity hybrid visual encoding framework for multimodal large language model. In *Advances in Neural Information Processing Systems*, volume 37, pages 101992–102010. Curran Associates, Inc.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024b. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, Miami, Florida, USA. Association for Computational Linguistics.

Jiri Martinek, Pavel Kral, Ladislav Lenc, and Josef Baloun. 2024. COMICORDA: Dialogue act recognition in comic books. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3566–3578, Torino, Italia. ELRA and ICCL.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *Preprint*, arXiv:2408.02718.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2025-05-18.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Reshma Ramaprasad. 2023. Comics for everyone: Generating accessible text descriptions for comic strips. *Preprint*, arXiv:2310.00698.

Christophe Rigaud, Jean-Christophe Burie, and Samuel Petit. 2024. Toward accessible comics for blind and low vision readers. *Preprint*, arXiv:2407.08248.

Ragav Sachdeva, Gyungin Shin, and Andrew Zisserman. 2024. Tails tell tales: Chapter-wide manga transcriptions with character names. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2053–2069.

Ragav Sachdeva and Andrew Zisserman. 2024. The manga whisperer: Automatically generating transcriptions for comics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12967–12976.

Ragav Sachdeva and Andrew Zisserman. 2025. From panels to prose: Generating literary narratives from comics. *Preprint*, arXiv:2503.23344.

Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24316–24342, Vienna, Austria. Association for Computational Linguistics.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, Wei Shi, Yuliang Liu, Hao Liu, Yuan Xie, Xiang Bai, and Can Huang. 2025. Textsquare: Scaling up text-centric visual instruction tuning. *Preprint*, arXiv:2404.12803.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, and 99 others. 2025a. Gemini robotics: Bringing ai into the physical world. *Preprint*, arXiv:2503.20020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 75 others. 2025c. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Yunjie Tian, Qixiang Ye, and David Doermann. 2025. Yolov12: Attention-centric real-time object detectors. *Preprint*, arXiv:2502.12524.

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

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024a. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *Preprint*, arXiv:2401.10529.

Yuhao Wang, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024b. MM-SAP: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9192–9205, Bangkok, Thailand. Association for Computational Linguistics.

Jianzong Wu, Chao Tang, Jingbo Wang, Yanhong Zeng, Xiangtai Li, and Yunhai Tong. 2025. Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation. *Preprint*, arXiv:2412.07589.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. 2024. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3783–3795.

xAI. 2025. Grok 3 beta — the age of reasoning agents. `https://x.ai/news/grok-3`. Accessed: 2025-05-04.

Min-Hsuan Yeh, Vincent Chen, Ting-Hao Huang, and Lun-Wei Ku. 2022. Multi-VQG: Generating engaging questions for multiple images. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 277–290, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lei Zhang, Fangxun Shu, Tianyang Liu, Sucheng Ren, Hao Jiang, and Cihang Xie. 2024a. Filter & align: Leveraging human knowledge to curate image-text data. *Preprint*, arXiv:2312.06726.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. 2024b. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19228–19252, Miami, Florida, USA. Association for Computational Linguistics.

A Annotation details

A.1 Detailes of Panel Segmentation and Ordering

We observed through extensive experimentation that existing multimodal large models exhibit significant limitations in comic panel recognition. Common issues include incomplete detection of speech balloons and misidentification of cover art as panels, leading to poor generalization across stylistically diverse comics.

To address these shortcomings, we curated a representative set of comic images from the GoComics website, annotating panels at a fine-grained level. We also incorporated two additional datasets: MangaZero (Wu et al., 2025) and COMICORDA (Martinek et al., 2024), forming a comprehensive training set that covers diverse comic styles and complex cases.

Using this merged dataset, we trained a state-of-the-art object detection model, YOLOv12 (Tian et al., 2025), which achieved strong generalization across different comic styles and accurately segmented panel regions within comic pages.

Following the implementation details from Sachdeva and Zisserman (2024), we represented comic layouts as Directed Acyclic Graphs (DAGs) and utilized topological sorting to establish the global reading order. Recognizing the distinct differences in reading conventions, we designed two separate rule sets to handle Western (left-to-right) and Japanese (right-to-left, top-right to bottom-left) comics. This pipeline facilitates precise detection and ordering of panels, providing a robust structural foundation for downstream narrative understanding and reasoning tasks.

A.2 Details of Image Preprocessing

We adopted different preprocessing strategies based on the type of comic images. For Western comics, which typically follow a left-to-right reading order, we first applied our panel segmentation and ordering pipeline described in the main text to identify each panel and assign it a sequential index. We then annotated the original comic page by drawing red bounding boxes around each panel and labeling them with the corresponding indices.

In contrast, Japanese comics—characterized by a right-to-left reading order, more complex layouts, and denser panel arrangements—posed greater challenges. During data generation, we found that adding visual prompts directly to full-page images did not significantly improve the model's ability to understand panel relationships. The model often struggled to capture the logical structure among panels. To address this, we cropped each panel as an individual image and sorted the resulting sequence according to the correct reading order. This ordered panel sequence was then used as input to the model.

A.3 Details of Description Annotation

Given the persistent challenges large-scale models face in accurately identifying comic panels and their reading order, we integrated our image preprocessing techniques with specifically designed prompts to guide model-based generation of structured textual descriptions.

For Western comics, we employed GPT-4o as the generation model; for Japanese comics, we used Qwen-Max. The input to each model consisted of the processed images (outputs from the previous stage) along with customized prompts tailored to the specific comic style. Each model produced a structured textual description segmented according to individual panels.

Following model generation, trained human annotators conducted comprehensive reviews of the generated descriptions. They corrected errors, supplemented missing details, and removed redundancies, ensuring each panel description was accurate, coherent, and readable.

A.4 Analysis of the Six Chronological Reasoning Perspectives

The six perspectives capture complementary aspects of chronological reasoning in comics. **Cross-panel action tracing** targets visually continuous actions (e.g., a thrown ball's trajectory), testing whether models can align events across panels. **Chronological character action sequences** assess whether actions by the same character form a coherent timeline, even without visual continuity, requiring causal inference. **Inferred motivations between non-consecutive panels** involve linking distant panels by understanding narrative intent. **Implicit character motives** demand recognizing subtle cues—like expressions or dialogue—to infer intent. **Prediction beyond shown panels** evaluates the model's ability to anticipate next events, while **counterfactual reasoning** asks it to imagine alternate outcomes.

B Experiment details

B.1 Evaluation Metric Definitions

To comprehensively evaluate model performance, we employ the following four metrics:

Accuracy Measures the proportion of panels placed in exactly the correct position:

$$\text{Accuracy} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{1}\big(p_i = c_i\big), \quad (5)$$

where M is the number of samples, N_j is the total number of panels in the j-th sample, p_i and c_i are the predicted and true positions of the ith panel, and $\mathbf{1}(\cdot)$ is the indicator function.

Mean Absolute Error (MAE) Computes the average absolute deviation between predicted and true positions:

$$\text{MAE} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}|p_i - c_i|, \quad (6)$$

where M is the number of samples, N_j is the total number of panels in the j-th sample, p_i is the predicted position of the i-th panel, and c_i is its true position.

Spearman Rank Correlation Applies the Spearman rank correlation coefficient to the predicted and true rankings to assess overall ordering consistency:

$$\rho = \frac{1}{M}\sum_{j=1}^{M}(1 - \frac{6\sum_{i=1}^{N_j}(r_i - s_i)^2}{N_j\,(N_j^2 - 1)}), \quad (7)$$

where r_i and s_i are the ranks of the ith element in the predicted and true orders, respectively. This can also be computed via `scipy.stats.spearmanr`.

NDCG@n Normalized Discounted Cumulative Gain at rank n, which emphasizes the quality of the top n panels:

$$\text{DCG}_n = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{n}\frac{\text{rel}_i}{\log_2(i+1)}, \quad (8)$$

$$\text{IDCG}_n = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{n}\frac{\text{rel}_i^{\text{ideal}}}{\log_2(i+1)}, \quad (9)$$

$$\text{NDCG}_n = \frac{1}{M}\sum_{j=1}^{M}\frac{\sum_{i=1}^{n}\frac{\text{rel}_i}{\log_2(i+1)}}{\sum_{i=1}^{n}\frac{\text{rel}_i^{\text{ideal}}}{\log_2(i+1)}}, \quad (10)$$

Model	Txt-only	Multimodal
GPT-4o	0.304	0.507
Moonshot-V1	0.328	0.396
Gemini-2.0	0.322	0.482
Qwen-Max	0.363	0.397
Grok-3	0.360	0.448
Gemma-3 (27B)	0.268	0.408
LLaMA-3.2-VI (90B)	0.255	0.438
Qwen-2.5-VL (32B)	0.316	0.382

Table 11: Results on *Multiple-Choice Question Answering* (Task 3) under text-only and full multimodal settings. **Txt-only** indicates that visual inputs are removed, while **Multimodal** denotes the default setting with both visual and textual inputs.

where rel_i is the relevance score of the ith panel in the predicted order, $\text{rel}_i^{\text{ideal}}$ is its score in the ideal order, and in our experiment, n is set to the actual number of panels on each page.

Accuracy directly measures exact panel placements. MAE, Spearman, and NDCG@n complement it by assessing error magnitude, ordering consistency, and top-n ranking quality for a more complete evaluation.

B.2 Additional Analysis on Multimodal Necessity

Table 11 compares model performance on the filtered MCQA subset under text-only and full multimodal settings, showing consistent improvements when visual inputs are available.

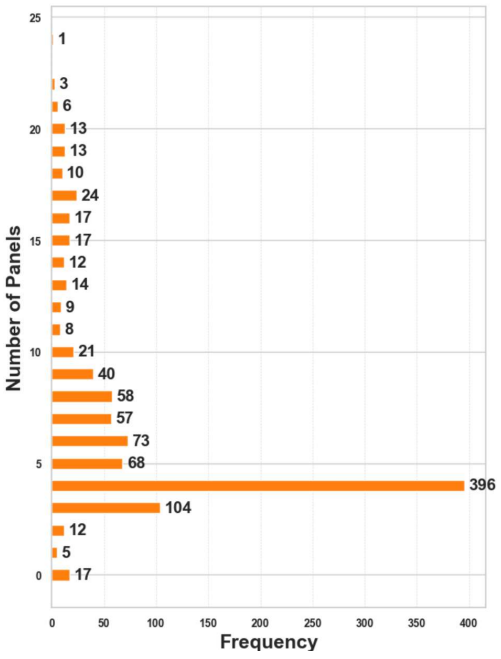


Figure 5: Distribution of panel counts in comic images

C Dataset Statistics

Figure 5 shows the distribution of panel counts in the CHROMIC dataset.

D Guidelines for Human Scoring of Descriptions

D.1 Task Objective

Given the **original comic panels** (with and without text) and the corresponding **textual description**, annotators are asked to judge how faithfully the description reflects the visual content of the panels.

D.2 Instructions

1. **Read the panels:** Observe character number, gender/appearance, main actions, scene elements, dialogue theme, and Chronological cues.
2. **Read the description:** Check whether it correctly captures the above elements.
3. **Compare and score:** Assign a consistency score using the 1–5 scale below.
4. **Scope:** Each description is scored once per comic instance.
5. **Independent annotation:** You will not know the text source, nor discuss with other annotators.

D.3 Scoring Scale (1–5)

- **5 = Highly Consistent**: Almost fully aligned with the panels; only minor omissions or wording differences.
  Example: Image: "boy in red shirt throws a ball"; Description: "a boy throws a ball".
- **4 = Mostly Consistent**: Core characters and main actions correct, but some secondary details missing.
  Example: Image: "two boys passing a ball"; Description: "a boy throws a ball".
- **3 = Roughly Consistent**: Partial alignment, but noticeable errors (e.g., role confusion).
  Example: Image: "boy throws ball to girl"; Description: "boy throws ball to a friend".
- **2 = Mostly Inconsistent**: Few correct elements, but main actions or roles are wrong.
  Example: Image: "boy throws a ball"; Description: "girl kicks a ball".
- **1 = Completely Inconsistent**: No meaningful alignment; content unrelated.
  Example: Image: "boy throws a ball"; Description: "a car parked on the street".

D.4 Additional Tag

Annotators may mark **"Uncertain"** if the panel is too ambiguous or the description too vague to judge reliably.

E Annotator Guidelines

E.1 Task Scope

Human annotators receive sentences or short paragraphs generated by the model and revise factual errors based on the reference comic image. In addition, they perform fine-grained edits to improve the accuracy, clarity, and fluency of the language, while preserving the original meaning.

E.2 Core Principles

- Ensure that the number and order of comic panels are consistent.
- Ensure that the description contains no factual errors that conflict with the original image.
- Ensure that each multiple-choice question has one and only one correct answer.
- When designing multiple-choice questions, refer to as many panels as possible, and avoid overly trivial questions.

E.3 Editing Workflow

1. Check whether the panel order is correct.
2. Read the model-generated description.
3. Revise the description based on the comic panels, following the core principles.
4. Read the model-generated multiple-choice question and additional candidate options.
5. Revise and filter the multiple-choice question based on the comic panels and core principles.
6. Select the three most plausible distractors from the six candidate options.

F Prompts for Data Generation and Evaluation

Prompt Template for Comic Description Generation :

Act as a visual storytelling expert.

First, generate a detailed comic strip description that:
1. Contains continuous temporal progression
2. Features at least 2 characters with observable physical/emotional changes
3. Includes implicit causal relationships between panels
4. Contains at least one scene transition (time/location change)
5. Uses vivid action verbs and spatial prepositions

Format requirements:
- Separate each panel with "Panel X:", but put all the panel descriptions in one paragraph, one line ("Panel 1: ... Panel 2: ... Panel N: ... ").
- When quoting a quote from the comic, use single quotes(') instead of double quotes("")
- Describe character positions/expressions in relation to objects
- Note subtle environmental changes between panels
- Avoid explicit explanations of character thoughts

Output Format
{
"description": "panel1 :... ; panel2 :... ; panel3 :... ; ... ; panelx :... ; "
}

Figure 6: This prompt template is designed to guide a large language model in generating structured and coherent comic panel descriptions.

Prompt for Comic Multiple-Choice Question Generation (A)

[DESCRIPTION]

Based on the comic description, create 4 challenging QA pairs focusing on:
A) Character action sequences requiring temporal reconstruction
B) Inferred motivations between non-consecutive panels
C) Prediction beyond shown panels (what happens next?)
D) Counterfactual reasoning (what if X panel changed?)

For each question:
1. Make answers require panel cross-referencing
2. Include at least one red herring detail from the text
3. Phrase 40% questions with negative premises ("Which did NOT...")
4. Provide detailed answer explanations identifying required reasoning types

Figure 7: This prompt template directs the model to generate four challenging multiple-choice QA pairs from a given comic description.

Prompt for Comic Multiple-Choice Question Generation (B)

[DESCRIPTION]

Based on the comic description, create 5 challenging QA pairs focusing on:
A) Cross-panel action tracing
B) Inferred motivations between non-consecutive panels
C) Implicit character motives
D) Prediction beyond shown panels (what happens next?)
E) Counterfactual reasoning (what if X panel changed?)

For each question:
1. Make answers require panel cross-referencing
2. Include at least one red herring detail from the text
3. Phrase 40% questions with negative premises ("Which did NOT...")
4. Provide detailed answer explanations identifying required reasoning types
5. Requires piecing clues from ≥3 panels
6. The question should take enough details into account and should be difficult enough and even GPT itself can hardly resolve it.

Figure 8: This prompt template directs the model to generate five challenging multiple-choice QA pairs from a given comic description.

Prompt for Multiple-Choice Distractor Generation

[DESCRIPTION]
[SELECTED QA-PAIR]

For the validated QA pair, create 3 distractors that:
1. Use logical fallacies common in temporal reasoning:
  - Post hoc ergo propter hoc (false causation)
  - Reverse chronology errors
  - Positional transposition
2. Incorporate actual elements from different panels
3. Match the grammatical structure of the correct answer
4. Include one "near miss" distractor requiring precise panel sequence recall
---
For each distractor:
- Identify which comic element it draws from
- Explain the cognitive bias it exploits
- Rate its deception level (1-5)

Figure 9: This prompt template guides the model to generate three high-quality distractors for a given, validated QA pair from a comic description.

Description Reordering Test Prompt

[IMAGE]

The following are shuffled descriptions of comic panels. Each line corresponds to one panel.

[0]. [Panel description 0]
[1]. [Panel description 1]
…
[N-1]. [Panel description N-1]

Now, return a Python list of integers indicating the correct order of these panels (starting from 0).
The list must contain exactly [N] elements—no more, no less—and must include every index shown above, without duplicates or omissions.

Do NOT return an empty list.
Return only the list of numbers, like this: [3, 0, 2, 1, 4]
Make sure the list includes all [N] panel indices.

Figure 10: This prompt template is designed to test a model's ability to reconstruct the original panel sequence from shuffled descriptions.

MCQ Answering Test Prompt

[IMAGE]

Question: [The multiple-choice question]
Options:
A. [option A]
B. [option B]
C. [option C]
D. [option D]

Just reply with a single letter.

Figure 11: This prompt template evaluates a model's ability to answer a comic-based multiple-choice question.