ComicVQA: A Benchmark for Visual Reasoning in Multimodal LLMs

Anonymous ACL submission

Abstract

We introduce Comic Visual Question Answering (**ComicVQA**), a comics-based benchmark for evaluating MLLMs on visual reasoning. ComicVQA comprises of (i) **Missing Panel Prediction**, testing fine-grained visual grounding and (ii) **Panel Sorting**, which evaluates sequential narrative understanding. Closed-source models achieve up to 70.4% on Missing Panel Prediction and 34.8% on Panel Sorting, whereas open-source models reach only 36.7% and 26.9%, respectively. With supervised fine-tuning on ComicVQA, Qwen2.5-VL-7B can close much of this performance gap, even surpassing GPT4o-mini on both tasks. Additionally, encouraging step-by-step reasoning over panel pairs can further improve Panel Sorting performance, suggesting avenues for aligning model reasoning with established chain-of-thought strategies. These results establish ComicVQA as a comprehensive benchmark for advancing visual reasoning in MLLMs.

1 Introduction

Robust visual reasoning requires both fine-grained visual grounding and, in some domains, the ability to integrate information across image sequences. Visual Question Answering (VQA) benchmarks such as VQA (Antol et al., 2015), GQA (Hudson and Manning, 2019) OK-VQA (Marino et al., 2019), and CLEVR (Johnson et al., 2017) have facilitated advances in models capable of integrating visual and textual information.

However, these datasets predominantly involve isolated single static images and textual answers, which allows models to exploit linguistic correlations rather than performing true visual grounding (Goyal et al., 2017; Agrawal et al., 2018; Liu et al., 2024). As a result, strong performance on existing VQA-style tasks does not necessarily imply robust visual reasoning, particularly in settings that require either fine-grained discrimination of visual details or integration of information across multiple images (Zhang et al., 2024; Hao et al., 2025; Xu et al., 2025). Although some recent work has begun to explore reasoning over multiple images or temporal visual streams (Xu et al., 2025; Imam et al., 2025) and narrative comics (Vivoli et al., 2025; Ikuta et al., 2025), comprehensive evaluation of fine-grained and sequential visual reasoning together remains limited.

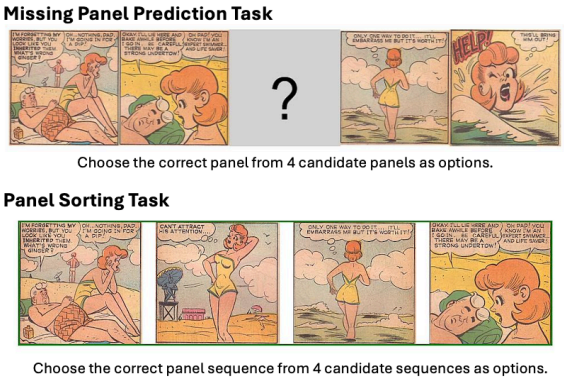

Figure 1: Snippets of our ComicVQA dataset, spanning over 2 tasks. All tasks are presented as multiple-choice questions, whereby the 4 options are all comic panels.

Comics provide a naturally rich domain for such evaluation. While they consist of images, these images are sequential and narrative driven. Understanding comics requires both fine-grained visual grounding such as recognizing characters, maintaining style continuity and resolving occlusions, and sequential reasoning to reconstruct the story (Iyyer et al., 2017; Vivoli et al., 2025). This makes comics an ideal setting to evaluate MLLMs on complementary aspects of visual reasoning that go beyond single image VQA (Vivoli et al., 2024).

To address this gap, we introduce Comic Visual Question Answering (**ComicVQA**), a benchmark designed to evaluate MLLMs on both fine-grained visual grounding and sequential narrative reasoning. ComicVQA contains two complementary multiple-choice tasks. The first task, *Missing Panel Pre-*

Feature / Dataset	COMICS	MangaUB	ComicsPAP	ComicVQA (Ours)
Sequential visual narratives	✓	✓	✓	✓
MCQs with image-only options	✗	✓	✓	✓
Missing panel prediction task	✗	✗	✓	✓
Panel metadata	✓	✗	✗	✓
Panel sorting task	✗	✗	✗	✓
Textually ambiguous distractors	✗	✗	✗	✓

Table 1: Comparison of ComicVQA with related benchmarks in the comics domain (Iyyer et al., 2017; Ikuta et al., 2025; Vivoli et al., 2025). ✓ indicates the dataset includes the feature; ✗ indicates it does not.

diction, targets fine-grained visual grounding by presenting candidate panels that are visually distinct but described with highly similar text, making textual cues insufficient for disambiguation. Models must therefore rely on visual features such as character appearance, artistic style, and continuity to select the correct panel. The second task, *Panel Sorting*, evaluates sequential reasoning by asking models to reconstruct the correct narrative order of panels. All answer options are images, which minimizes the influence of language priors and compels models to perform direct visual comparison.

We evaluate a range of open-source and proprietary MLLMs on both ComicVQA tasks. On Missing Panel Prediction, open-source models achieve up to 36.7% accuracy, while proprietary models reach 70.4%. On Panel Sorting, open-source and proprietary models reach up to 26.9% and 34.8%, respectively, close to random scoring. For Missing Panel Prediction, most models perform consistently better on the image-based version than on a text-only counterpart generated from GPT4o-mini panel descriptions. These results indicate that ComicVQA encourages genuine visual reasoning for fine-grained understanding while posing a challenging benchmark for sequential reasoning.

By providing visually discriminable but semantically similar answer choices for the Missing Panel Prediction task, and a complementary Panel Sorting task that probes sequential narrative understanding, ComicVQA establishes a rigorous testbed for advancing multimodal visual reasoning. Our contributions are: (i) the ComicVQA dataset and task suite, which combines Missing Panel Prediction and Panel Sorting to probe fine-grained grounding and sequential understanding; (ii) a comprehensive evaluation of open-source and proprietary MLLMs across image and text modalities; (iii) we will be releasing the dataset, evaluation code, and annotations to support further research. Snippets of our ComicVQA dataset is as shown in Figure 1.

2 Related Work

Visual Question Answering. Visual Question Answering (VQA) has been central to multimodal reasoning, with benchmarks such as VQA (Antol et al., 2015), CLEVR (Johnson et al., 2017), GQA (Hudson and Manning, 2019), and OK-VQA (Marino et al., 2019). While these datasets drove progress, they mainly text-based answers, leaving models prone to language priors rather than true visual grounding (Goyal et al., 2017; Agrawal et al., 2018; Liu et al., 2024). More recent benchmarks introduce pattern-recognition (Xu et al., 2025), chart-based (Masry et al., 2022), and mathematical tasks (Zhang et al., 2024). Multi-image reasoning efforts (Vivoli et al., 2025; Imam et al., 2025) emphasize context across frames, but focus on motion continuity rather than discrete static sequences.

Comics and Narrative Understanding. Comics offer a rich testbed for sequential and narrative reasoning. The COMICS dataset (Iyyer et al., 2017) introduced cloze-style QA, and subsequent works extended comics understanding to retrieval, segmentation, and narrative inference (Aizawa et al., 2020; Vivoli et al., 2024, 2025; Ikuta et al., 2025). However, most existing datasets emphasize dialogue recovery or mixed text–image reasoning rather than controlled, visual-only evaluation. ComicVQA fills this gap by designing tasks that reduce textual dependence and unify fine-grained visual grounding with sequential narrative understanding. Table 1 summarizes how our dataset extends prior comics or manga based benchmarks.

3 Methodology

3.1 Task Definition

We define two core visual QA tasks over multi-panel comic sequences, each formulated as a multiple-choice question with four options.

Task 1: Panel Sorting. Given four shuffled comic panels, the model must identify the correct

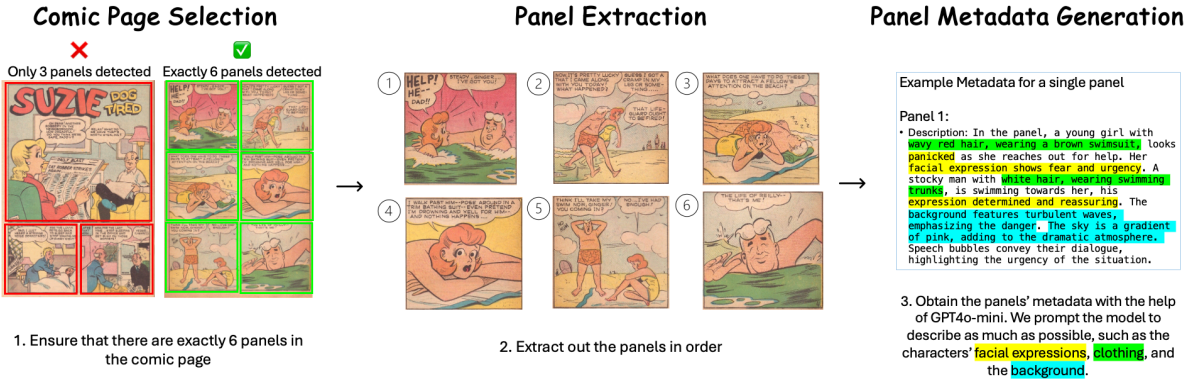

Figure 2: ComicVQA dataset construction consists of three steps: 1. Filtering down the comic pages to the ones with exactly six panels, 2. Extraction of panels from the finalized comic pages, and 3. Generate metadata for each panel for analysis.

narrative order from four candidate permutations. This task primarily assesses the model's ability to capture temporal progression and causal relationships, requiring coherent narrative understanding across sequential images.

Task 2: Missing Panel Prediction. Given an incomplete comic strip, the model must select the panel that correctly fills the masked panel. Unlike Panel Sorting, this task is designed to stress fine-grained visual discrimination rather than narrative inference. Candidate panels are deliberately chosen to be textually similar to the ground-truth panel, forcing the model to rely on perceptual cues such as character appearance or artistic style rather than language priors or superficial semantic overlap.

Each task's input to the model consists of a prompt text specifying the task and a composed question image. Outputs are free-form text responses, which can take the form of direct answers or additional reasoning strategies implemented depending on the prompt used and model capabilities.

3.2 Dataset

3.2.1 Dataset Construction

We construct our dataset from a subset of the COMICS dataset (Iyyer et al., 2017), which comprises comic panels sourced from the public-domain Digital Comics Museum (DCM)[1], a repository of Golden Age comic books that are no longer under copyright and are in the public domain. To complement this subset, we also self-sourced pages from DCM and segmented them using a panel segmentation algorithm inspired by Manga-Panel-Extractor[2]. For uniformity in input and question design, we select pages with exactly six detected panels. Although most questions involve four to five panels, retaining six preserves page structure. The selected panels are recomposed into composite images for various QA tasks, and all question images are resized to a maximum resolution of 1024×1024 for compatibility with open-source vision models.

In addition to the visual data, we include metadata for each panel: a 100-word description of a comic panel, generated using GPT-4o-mini (Hurst et al., 2024). We use the following prompt:

> Describe the panel's visual content thoroughly, including the characters' poses, facial expressions, actions, clothing, background, and any visual storytelling cues. Maximum 100 words.

The full pipeline is as illustrated in Figure 2.

3.2.2 Question Design

We present all panels within a single composite image rather than providing separate images. This design choice also circumvents limitations of certain MLLMs that may not support multi-image inputs. To facilitate understanding of the question image, we further annotate the composite image with visually distinct bounding boxes. Following the common practice in object detection tasks, whereby bounding boxes are used to highlight target regions (Girshick, 2015; Redmon et al., 2016), we employ red bounding boxes to denote the question context panels and green bounding boxes to mark the candidate answer panels. Each candidate option is

[1] https://digitalcomicmuseum.com/

[2] https://github.com/adenzu/Manga-Panel-Extractor

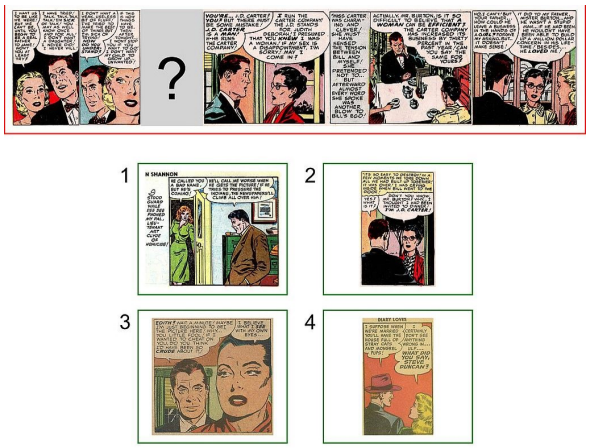

Figure 3: Example of a Missing Panel Prediction question. Across all candidate panels, they depict a man and a woman with multiple dialogue bubbles. The correct option (2) can be quickly identified through visual cues like the artistic style and characters design, without the need to read though the dialogue text to determine the correct narrative sequence.

additionally labeled with a numeric identifier, enabling the model to output a discrete option number as its prediction. Examples illustrating the question design for both tasks are provided in Figure 3 and Appendix D.1.

3.2.3 Dataset Manipulation

To increase the difficulty of the Missing Panel Prediction task, we introduce targeted manipulations. Instead of selecting candidate panels at random for multiple-choice questions, we leverage GPT4o-mini's generated descriptions to identify textually similar panels. We used embeddings generated by text-embedding-3-large (OpenAI, 2024) on each of the panel's descriptions, and compute cosine similarity with the solution panel to select the top three most similar candidates from the same split. This ensures that candidate panels are textually similar, making purely language-based reasoning insufficient and requiring models to use visual cues. An example of a question with visually distinct but textually similar candidates is shown in Figure 3.

3.2.4 Dataset Statistics

The ComicVQA dataset comprises a total of 6157 multiple-choice questions. Panel Sorting has 2,000 train, 500 valid, and 1,000 test questions, while Missing Panel Prediction has 1,146 train, 511 valid, and 1,000 test questions.

On average, each question includes 4 to 8 unique panels, depending on the task structure. This design aims to have a unified input format across all questions. All textual content within comic panels, including narration and speech bubbles, are provided in English only. While the options were randomly shuffled, we ensured that all answer positions were approximately balanced, resulting in an equal distribution across the dataset. Additionally, for the Missing Panel Prediction task, we ensure that the masked panels are evenly distributed across all positions across the panel sequence.

3.3 Prompt Settings

We evaluate model performance using the zero-shot prompting approach (Kojima et al., 2022) and provide guidance highlighting context panels in red and candidate panels in green. Below, we show the prompts used for each task.

- **Panel Sorting.** Q: Four different sequences of the same set of comic panels are shown, labeled Options 1, 2, 3, 4. Each option shows a different panel order. Which of the 4 green-boxed panels below is the correct panel sequence? Choose the correct option. Please conclude your answer with 'The answer is: Option (?)'. A:

- **Missing Panel.** Q: The comic panels in the red box show the story with one missing panel labeled '?'. Which of the 4 green-boxed panels below is the correct missing panel? Options 1 and 2 are in the first row, and Options 3 and 4 are in the second row. Choose the correct option. Please conclude your answer with 'The answer is: Option (?)'. A:

4 Experiment

4.1 Experimental Setup

Dataset. All evaluations are conducted on the test split of the ComicVQA dataset, which consists of 1000 multiple-choice questions per task, totaling 2000 test instances.

Models. We evaluate a range of open-source and proprietary MLLMs. We used LLaVA-1.6-Mistral-7B (Liu et al., 2023; Jiang et al., 2023), Qwen2.5-VL-7B-Instruct (Wang et al., 2024; Bai et al., 2023, 2025), LLaVa-1.6-Vicuna-13B (Liu et al., 2023; Chiang et al., 2023), Llama-3.2-11B-Vision-Instruct (AI, 2024), Gemma-3-27b-it (Team et al., 2025), GPT4o-mini (Hurst et al., 2024), and GPT4.1-mini (OpenAI, 2025). All models are instruction tuned models.

Category	Model (#Params)	Missing Panel (%)	Panel Sorting (%)
Baseline	Random	25.0	25.0
Open-Source	LLaVA-v1.6-Mistral-7B (7.57B)	24.8	25.1
	Qwen2.5-VL-7B-Instruct (8.29B)	<u>36.7</u>	26.8
	LLaMA-3.2-Vision-Instruct (10.7B)	25.3	26.6
	LLaVA-v1.6-Vicuna-13B (13.4B)	24.6	25.4
	Gemma-3-27b-it (27.4B)	33.5	<u>26.9</u>
Proprietary	GPT4o-mini	61.6	24.7
	GPT4.1-mini	<u>70.4</u>	<u>34.8</u>
Finetuned	Finetuned-MPP Qwen2.5-VL-7B-Instruct (8.29B)	62.1	-
	Finetuned-PS Qwen2.5-VL-7B-Instruct (8.29B)	-	28.4

Table 2: Accuracy (%) of instruction-tuned models on the two ComicVQA tasks. "Finetuned-MPP" indicates models fine-tuned on the Missing Panel Prediction training set, and "Finetuned-PS" indicates models fine-tuned on the Panel Sorting training set. Highest accuracies under each category are underlined.

Implementation details. We present accuracy results for both tasks. All model evaluations were conducted on H100 GPUs with 80GB of VRAM. Model hyperparameters and model evaluation information are indicted in detail in Appendix A.

4.2 Main Results

Table 2 presents the performance of all evaluated models on ComicVQA's Missing Panel Prediction and Panel Sorting tasks. Random guessing provides a baseline accuracy of 25% for both tasks.

Missing Panel Prediction. Performance on this task varies widely across models. Smaller open-source models such as LLaVA-v1.6-Mistral-7B (24.8%), Llama3.2-Vision-Instruct (25.3%), and LLaVA-Vicuna-13B (24.6%) achieve near-random accuracy, while Qwen2.5-VL-7B-Instruct improves to 36.7%, indicating a better ability to leverage visual cues. Proprietary models such as GPT-4o-mini (61.6%) and GPT-4.1-mini (70.4%) achieve the highest accuracies, reflecting strong multimodal reasoning. Notably, Gemma-3-27B-it attains only 33.5%, performing worse than the much smaller Qwen2.5-VL-7B-Instruct model, indicating that model scale alone does not ensure effective visual reasoning.

Panel Sorting. This task, which evaluates temporal and causal reasoning, remains substantially more challenging than Missing Panel Prediction. With the exception of GPT4.1-mini (34.8%), most models perform near the random baseline of 25%, with accuracies ranging from 24.7–26.9%. Gemma-3-27B-it (26.9%), LLaMA-3.2-Vision-Instruct (26.6%), and Qwen2.5-VL-7B-Instruct (26.8%) achieve the best open-source results, while GPT4o-mini (24.7%) and smaller LLaVA variants (25.1–25.4%) remain close to random guessing. These findings highlight that narrative ordering is a difficult open problem for current multimodal models, where improvements depend less on scale and more on the capacity to integrate sequential visual context. This underscores ComicVQA's value in assessing higher-level narrative reasoning, a dimension rarely addressed in existing benchmarks.

Overall, these results suggest that visual reasoning for Missing Panel Prediction is both achievable and highly sensitive to model architecture and multimodal capabilities, whereas narrative ordering as tested in Panel Sorting remains more challenging for current vision-language models.

4.3 Fine-tuning

Additionally, we conducted a simple LoRA fine-tuning experiment on the Qwen2.5-VL-7B-Instruct model to evaluate the utility of our dataset. As shown in Table 2, fine-tuning on our training set enables Qwen2.5-VL-7B-Instruct to reach 62.1% and 28.4% accuracy respectively, outperforming the much larger model GPT4o-mini for both tasks. The substantial performance gain from lightweight adaptation suggests that our dataset provides novel, domain-specific supervision that can be readily leveraged to improve multimodal reasoning. The finetuning details and hyperparameters can be found in Appendix B.

4.4 Ablation: Impact of Visual Information

To isolate the role of visual understanding, we conduct an ablation study where all image inputs are replaced with textual descriptions for the Missing Panel Prediction task. These are generated by GPT4o-mini, which provides details on each panel's visual content, including characters, ac-

Models(#Params)	Image	Text
LLaVA-v1.6-Mistral-7B (7.57B)	24.8	25.4
Qwen2.5-VL-7B-Instruct (8.29B)	36.7	35.9
LLaMA-3.2-Vision-Instruct (10.7B)	25.3	40.4
LLaVA-v1.6-Vicuna-13B (13.4B)	24.6	24.5
Gemma-3-27b-it (27.4B)	33.5	46.8
GPT-4o-mini (Proprietary)	61.6	28.5
GPT-4.1-mini (Proprietary)	70.4	53.1

Table 3: Accuracy (%) of models on the **Missing Panel Prediction** task using the original image-based questions versus the text-only conversion. This highlights the performance changes when visual cues are removed.

Comparison	Perturbation Type	Accuracy (%)
12 vs **21**	Adjacent (1 ↔ 2)	60.1
23 vs **32**	Adjacent (2 ↔ 3)	59.4
34 vs **43**	Adjacent (3 ↔ 4)	57.8

Table 4: Diagnostic accuracy of 2 panels across different perturbation settings, while keeping 2 panels correct. The model detects the front 2 panels better than the other panel pairs.

Comparison	Perturbation Type	Accuracy (%)
1234 vs **21**34	Adjacent (1 ↔ 2)	57.4
1**23**4 vs 1**32**4	Adjacent (2 ↔ 3)	57.2
12**34** vs 12**43**	Adjacent (3 ↔ 4)	54.2
123**4** vs **4**23**1**	Global	61.3

Table 5: Diagnostic accuracy of 4 panels across different perturbation settings, while keeping 2 panels correct. The model detects large global disruptions more reliably than subtle local swaps.

tions, and scene context. This representation preserves fine-grained cues relevant for grounding while removing raw visual information, allowing us to evaluate whether models can solve the tasks using text alone. An example of such a textual description is shown in Figure 2.

When querying the model, we keep the prompt as close as possible in wording to the original image-based task prompt, removing only words that are specific to visual inputs.

Table 3 compares model performance between the image-only and text-only versions of the Missing Panel Prediction task. Four out of seven models perform better with visual inputs, indicating that image-based cues do help to contribute to task performance. GPT4o-mini achieves 61.6% accuracy with images compared to 28.5% with text, while GPT4.1-mini's performance went up from 53.1% to 70.4% when the question is converted from text to image. Among open-source models, Qwen2.5-VL-7B-Instruct (36.7% → 35.9%) and LLaVA variants exhibit comparable performance across modalities, suggesting that textual descriptions do not provide additional benefits beyond visual information. In contrast, Gemma-3-27B-it improves from 33.5% (image) to 46.8% (text), and LLaMA-3.2-Vision-Instruct from 25.3% to 40.4%, implying that some models still rely on linguistic priors. Overall, these results demonstrate that ComicVQA encourages visual reasoning in MLLMs, as models must leverage visual features to perform well on the task.

5 Analysis of Model Performance

5.1 Breakdown Analysis on Panel Sorting

Overall performance on the panel sorting task is close to random for all models, with the exception of GPT4.1-mini, which performs better at 34.8%. To understand where current models struggle and how future models might better handle large-image sequencing, we analyze GPT4.1-mini in detail, being the highest performing model.

Two-panel comparisons. Panel sorting can be viewed as a sequence of local comparisons between adjacent pairs. Based on this intuition, we evaluate the model's accuracy when restricted to ordering neighboring panels. Starting from sequences of four panels, we grouped them into two consecutive pairs and introduced controlled perturbations, reducing the search space from four possible configurations to two.

As shown in Table 4, the model achieves accuracies of 60.1%, 59.4%, and 57.8% for the (1↔2), (2↔3), and (3↔4) swaps respectively, all above the 50% random guessing score. This indicates that the model is able to capture local coherence when the comparison space is limited, although its performance decreases for later panel positions. These results indicate that the difficulty of the full four-panel sorting task may stem from the need to integrate information across multiple panels simultaneously. An example of a two-panel comparison question is provided in Appendix E.1.

Four-panel comparisons. Given the two-panel performance, we then expand the analysis to four-panel sequences, fixing two panels and permuting the remaining two. This reduces the task to a two-option comparison, similar to the two-panel case, but with additional surrounding context. Overall

(a) Image Input

(b) Text Input

Figure 4: Word clouds for GPT4o-mini on the Missing Panel Prediction task when given image input vs. text input.

accuracy is slightly lower or comparable to the two-panel condition, reflecting the increased difficulty introduced by additional contextual dependencies. As shown in Table 5, accuracy for swapping the first two, middle two, and last two panels are 57.4%, 57.2%, and 54.2%, respectively. Notably, performance peaks at 61.3% for the global swap (1↔4), suggesting that the model more easily distinguishes large narrative disruptions than subtle differences between adjacent panels.

Amongst the adjacent perturbations, the easiest case is swapping the first two panels (**12**34 vs. **21**34), reflecting the model's natural left-to-right processing bias. Swapping the last two panels (12**34** vs. 12**43**) is more difficult, as it requires integrating later context before resolving earlier panels. Reordering the middle two panels similarly poses a challenge, as the model must consider both preceding and following context simultaneously. These results suggest that GPT4.1-mini exhibits a strong sequential, left-to-right narrative bias, but can leverage additional context when disruptions are large or obvious. An example of a four-panel comparison question is provided in Appendix E.2.

Importantly, this pattern aligns with ComicVQA's objective of probing narrative reasoning: the model succeeds when coherence breaks are overt but struggles with local reordering that requires nuanced story comprehension.

Three-panel windows. Finally, we also evaluate model performance on windows of three consecutive panels, testing across four randomly generated permutations. Accuracy drops to 31%, lower than the four-panel setting. The relatively higher success on four-panel sequences (34.8%) appears to stem from clearer visual and narrative differences between the first and fourth panels as noted in Table 5. In contrast, the first and third panels in three-panel sequences might be closer in context and narrative, thus harder for the model to discriminate. An example of a three-panel comparison question is provided in Appendix E.3.

Adding instructions to the prompt. Taken together, these results suggest that models can perform reliable two-panel comparisons but struggle as the number of panels increases. This mirrors the strategy of performing local comparisons first and then composing them into a global ordering. Inspired by this observation, we modify the model prompting to explicitly encourage step-by-step reasoning through successive two-panel comparisons. The new prompt includes the following instructions in italics:

> Four different sequences of the same set of comic panels are shown, labeled Options 1, 2, 3, 4. Each option shows a different panel order. Which of the 4 green-boxed panels is the correct panel sequence? *For each option sequence, consider panels by pairs from left to right to determine the correctness of the story sequence.* Choose the correct option.

This added instruction improves performance by 5 percentage points, increasing accuracy to 39.8%, highlighting the benefit of aligning model reasoning with step-by-step incremental strategies. An example of a successful step-by-step incremental reasoning strategy is shown in Figure 5.

5.2 Qualitative Analysis of Model Outputs for Missing Panel Prediction

To compare the reasoning process of models when given images as compared to only text, we generate word clouds of the top used words by GPT4o-mini for the text-only and image versions of the

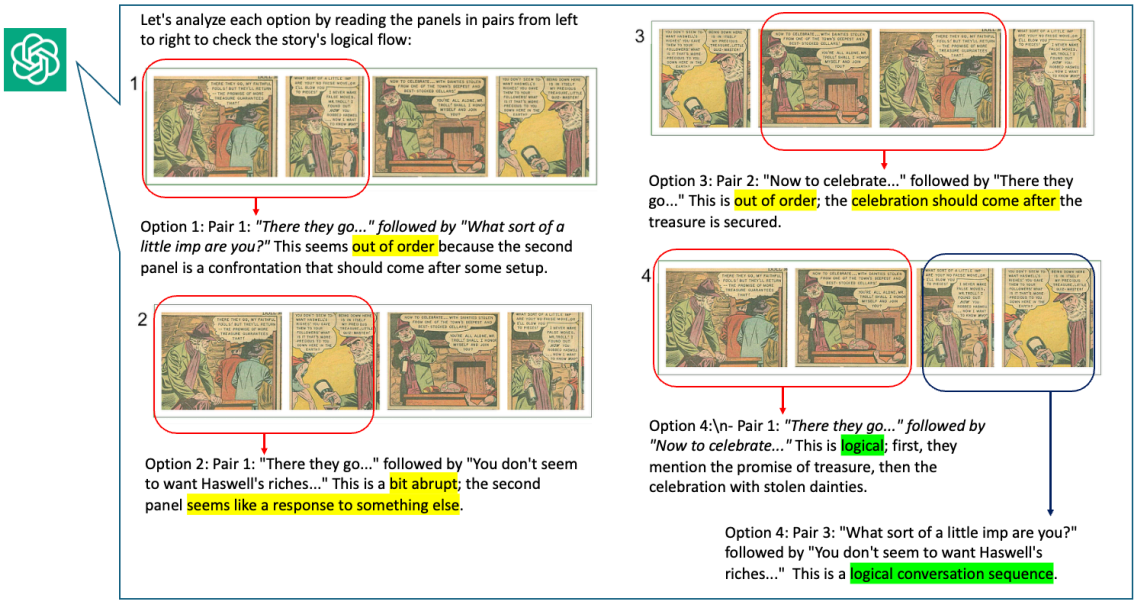

Figure 5: GPT4.1-mini's reasoning excerpts for an example question, showing how highlighted phrases indicate discrimination (yellow) and logical flow (green) when judging panel pairs. Main deciding statements are shown. We can see that GPT4.1-mini does pairwise comparison to determine if the option is correct or wrong.

Missing Panel Prediction task. We exclude stopwords and necessary task related words such as "option" and "panel". As shown in Figure 4a, when given images, models focus on visual features and words related to visual descriptions of the panels (e.g. "shows," "surrounding," "situation," and "dialogue"). In contrast, when given text, as shown in Figure 4b, models tend to regurgitate words from their prompts (e.g. "man, " "woman," "urgency," and "tension" which are all common words in the panel descriptions). This illustrates that models reason differently when using images than text; they are not simply performing the same reasoning after extracting information from the panels.

5.3 Summary of Analysis

Across both the panel sorting and missing panel prediction tasks, our results show that models continue to face challenges in visual reasoning. In the panel sorting experiments, models struggle on localized two-panel (adjacent) comparisons, as compared to global two-panel comparisons suggesting difficulty in fine-grained narrative discrimination. The qualitative word cloud analysis of the missing panel task further demonstrates that models process images and text differently, showing how the dataset encourages visual reasoning away from language priors. Together, these results illustrate that our benchmark provides a comprehensive framework for evaluating models' visual reasoning skills: it captures both the ability to track sequential narratives and the capacity to leverage visual information beyond textual cues.

6 Conclusion

In this work, we introduced ComicVQA, a benchmark for fine-grained visual grounding and narrative reasoning in comics, featuring two tasks: Missing Panel Prediction and Panel Sorting. Our evaluation of MLLMs highlights challenges across the two tasks: Missing Panel Prediction encourages visual-based reasoning, as most models perform poorly on text-only inputs, while Panel Sorting tests sequential narrative understanding, where even large models struggle. ComicVQA provides a platform for studying visual reasoning beyond simple image-text alignment and supports the development of models capable of deeper multimodal understanding. We hope it spurs further research in multimodal narrative reasoning and the creation of models that effectively interpret sequences of images in context. Future work includes expanding this work other visual-based domains that involve temporal and visual aspects (e.g., educational comics, video), and leveraging advanced multimodal pretraining to improve complex visual and contextual reasoning.

Limitations

One limitation of our benchmark is that it is currently restricted to English Language only, which may limit accessibility and model generalization across languages. Additionally, the comics used in our dataset are culturally situated, often reflecting Western storytelling conventions, humor, and norms that may not translate well to other regions or model pretraining distributions. Finally, all images are standardized to a resolution of a maximum of 1024×1024 pixels. While this enables uniform processing, it may pose challenges for models with limited vision capacity or without strong hierarchical visual encoders, particularly smaller architectures, which imply that they might not be looking at the full question image in the first place, hence not reflecting the true performance of the model.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

Meta AI. 2024. Lllama 3.2: Revolutionizing edge ai and vision with open, customizable models. Accessed: 2025-10-05.

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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

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. *arXiv preprint arXiv:2502.13923*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hikaru Ikuta, Leslie Wohler, and Kiyoharu Aizawa. 2025. Mangaub: A manga understanding benchmark for large multimodal models. *IEEE MultiMedia*.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. 2025. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *arXiv preprint arXiv:2501.10674*.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, 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*, pages 7186–7195.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744. Accessed: 2025-10-05.

Ying Liu, Ge Bai, Lu Chenji, Shilong Li, Zhang Zhang, Ruifang Liu, and Wenbin Guo. 2024. Eliminating the language bias for visual question answering with fine-grained causal intervention. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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

OpenAI. 2024. New embedding models and api updates. Accessed: 2025-10-05.

OpenAI. 2025. Introducing gpt-4.1 in the api. Accessed: 2025-10-05.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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.

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

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

A Model Hyperparameters

To ensure all results are reproducible and consistent, all models have the `temperature` parameter set to 0. Additionally, `do_sample` is set to False for open-sourced models, and `top_p` is set to 1 for OpenAI models. As for token generation, to ensure a fair comparison, the maximum number of tokens allowed to generate (`max_new_tokens`) is set to 2048 tokens only, to ensure that all models have the opportunity to explain the question throughly before answering the question. Smaller number of tokens have been attempted, but are often cut off especially for the text based questions in the ablations.

Additionally, because the outputs are deterministic with the above-mentioned settings, we only ran single-runs for all experiments.

Time taken to run a model on the test split maximally takes 5 hours, but factors such as the size of the model and the current load on the GPU can reduce the time taken for a single run on the test split to be less than 5 hours.

B LORA Finetuning

B.1 Hyperparameters used for Missing Panel Prediction Task

We fine-tuned the base model using Low-Rank Adaptation (LoRA) with a rank of 8, a scaling factor (lora_alpha) of 16, and a LoRA dropout rate of 0.05. Adaptation was applied only to the query and value projection layers (q_proj and v_proj), with no bias terms adapted. The LoRA task type was set to causal language modeling since we are using Qwen2.5-VL-7B-Instruct model. Training was conducted for 10 epochs with a per-device

batch size of 4 for both training and evaluation, and gradients were accumulated over 8 steps. We used the adamw_torch_fused optimizer with a constant learning rate of 2×10e-4, a maximum gradient norm of 0.3, and a warmup ratio of 0.03. Evaluation was performed every 10 steps, and models were saved every 20 steps. The best model was selected based on the lowest evaluation loss.

B.2 Hyperparameters used for Panel Sorting Task

For Panel Sorting Task, as there is a greater number of train data as compared to Missing Panel Prediction Task, we reduced the number of training epochs. We fine-tuned the base model using Low-Rank Adaptation (LoRA) with a rank of 8, a scaling factor (lora_alpha) of 16, and a LoRA dropout rate of 0.05. Adaptation was applied only to the query and value projection layers (q_proj and v_proj), with no bias terms adapted. The LoRA task type was set to causal language modeling since we are using Qwen2.5-VL-7B-Instruct model. Training was conducted for 3 epochs with a per-device batch size of 4 for both training and evaluation, and gradients were accumulated over 8 steps. We used the adamw_torch_fused optimizer with a constant learning rate of 2×10e-4, a maximum gradient norm of 0.3, and a warmup ratio of 0.03. Evaluation was performed every 10 steps, and models were saved every 20 steps. The best model was selected based on the lowest evaluation loss.

C Licenses

We obtained our data from 2 main areas: the COMICS dataset (Iyyer et al., 2017) and from Digital Comics Museum.

COMICS dataset can be used under the MIT License[3].

Digital Comics Museum is a publicly available data source for Golden Age comics.[4]

D Example Questions from Panel Sorting Task

D.1 Panel Sorting Example Question

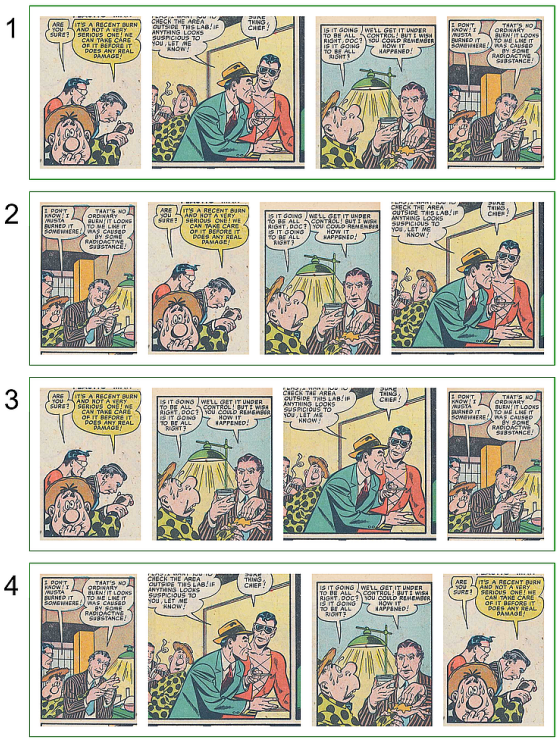

Figure 6: The panel sorting task provides 4 different permutations of a sequence of 4 comic panels. One of them is the correct sequence, and 3 others are shuffled.

[3]https://github.com/miyyer/comics/blob/master/LICENSE
[4]https://digitalcomicmuseum.com/forum/index.php

E Analysis Example Questions

E.1 Panel Sorting 2 Panel Example Question

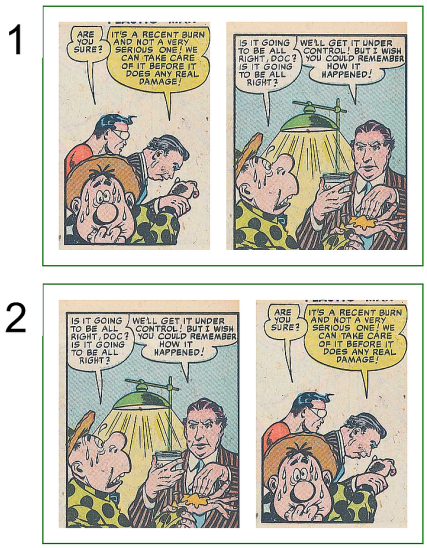

Figure 7: For analysis, we cut down the number of panels to see any changes in the performance of the model with a smaller number of candidate options. Since 2 panels only have 2 permutations, theres only 2 options.

E.2 Panel Sorting 4 Panel-2 Option Example Question

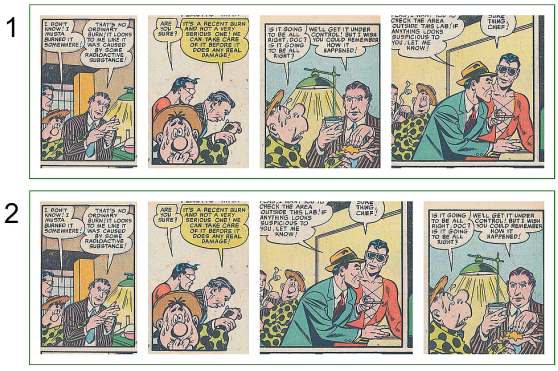

Figure 8: Simlar to the 2 panel analysis, we keep 2 panels in the correct order, and shuffled the remaining 2.

E.3 Panel Sorting 3 Panel Example Question

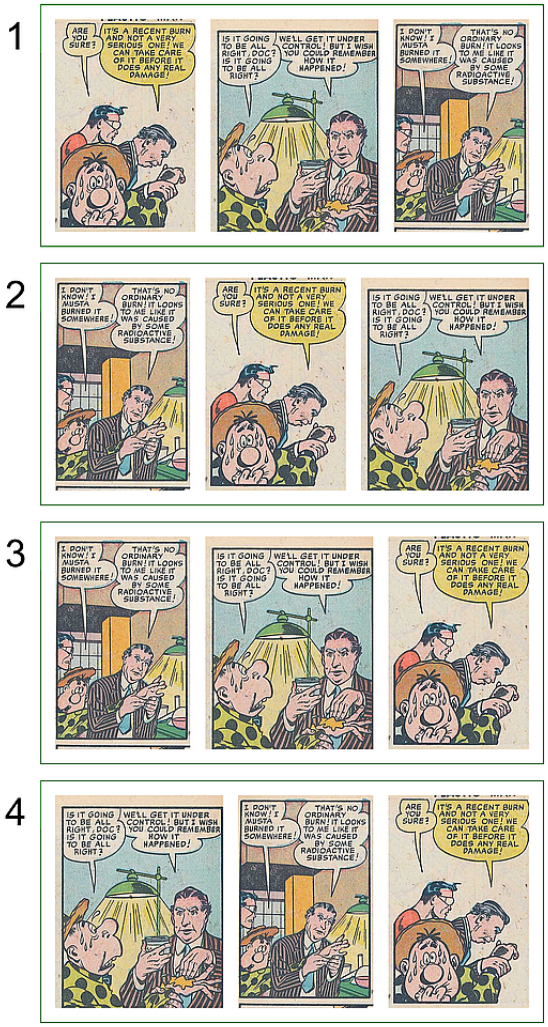

Figure 9: We also tried 3 panel permutations, since there are 6 possible permutations, we randomly selected 3 alongside the correct sequence.