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. Proprietary models achieve up to 62.6% on Missing Panel Prediction and 46.4% on Panel Sorting, whereas open-source models reach only 47.7% and 26.9%, respectively. In contrast, human annotators achieve over 83% accuracy on both tasks, revealing a large gap between current models and human-level multimodal understanding in comics. Through controlled ordering ablations and a detailed error taxonomy, we show that current MLLMs rely primarily on coarse temporal cues and struggle with fine-grained visual reasoning. These findings demonstrate ComicVQA as a diagnostic benchmark for advancing multimodal visual reasoning in comics.

1 Introduction

Robust visual reasoning requires both fine-grained visual grounding and, in many real-world settings, the ability to integrate information across sequences of images. 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 driven significant progress in models that process multimodal inputs. However, these benchmarks predominantly focus on isolated static images paired with short textual answers.

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). This limitation becomes particularly pronounced in settings that require either fine-grained discrimination of visual details or the integration of information across multiple images (Hu et al., 2024; Zhang et al., 2024; Hao et al., 2025; Xu et al., 2025). While recent efforts have begun to explore reasoning over multiple images, temporal visual streams, or structured visual contexts (Xu et al., 2025; Imam et al., 2025; Ryan et al., 2025), comprehensive evaluation frameworks that jointly probe fine-grained visual grounding and sequential visual reasoning remain limited.

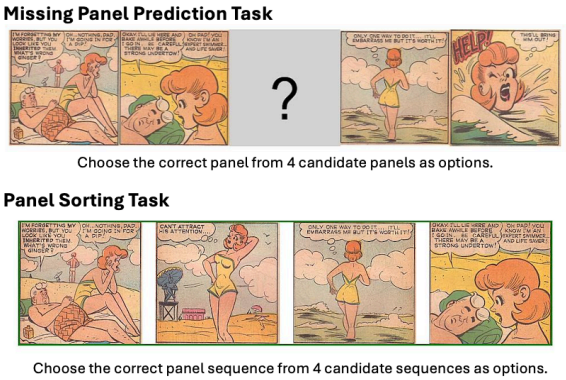

Figure 1: Snippets of our ComicVQA dataset, spanning over 2 tasks. All tasks are presented as multiple-choice questions, whereby the 4 options are all comic panels.

Comics provide a natural and underexplored domain for evaluating these complementary aspects of visual reasoning. Although comics consist of static images, they are inherently sequential and narrative-driven. Understanding a comic strip requires tracking characters and objects across panels, maintaining consistency of visual states, resolving scene transitions, and inferring causal and temporal relationships between events (Iyyer et al., 2017; Vivoli et al., 2025). Unlike videos, comics present sparse, discrete visual evidence, making narrative reconstruction particularly challenging and placing greater demands on visual continuity and inference. These properties make comics a compelling testbed

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

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

for evaluating multimodal visual reasoning beyond single-image VQA (Vivoli et al., 2024).

We introduce Comic Visual Question Answering (ComicVQA), a diagnostic benchmark designed to assess complementary reasoning skills in comics through two multiple-choice tasks (Figure 1):

Missing Panel Prediction. This task tests on fine-grained visual discrimination by removing all text found on the panels. Models select the correct panel using visual cues such as artistic style, character appearance, and background scenes.

Panel Sorting. This task evaluates compositional temporal reasoning by requiring models to determine the correct sequence of panels.

We evaluate a broad set of open-source and proprietary MLLMs on both tasks. On Missing Panel Prediction, open-source models achieve up to 47.7% accuracy, while proprietary models reach 62.6%. Panel Sorting proves substantially more challenging: open-source models perform near random (up to 26.9%), and the best proprietary model reaches only 46.4%. In contrast, human annotators exceed 83% accuracy on both tasks, revealing a substantial gap between current models and human-level multimodal visual understanding.

By combining visually discriminative but semantically similar answer choices in Missing Panel Prediction with a complementary Panel Sorting task that probes narrative reasoning, ComicVQA provides a rigorous testbed for advancing multimodal visual reasoning in comics. Our contributions are: (i) ComicVQA, a benchmark unifying fine-grained visual grounding and sequential narrative reasoning in comics; (ii) a comprehensive evaluation of MLLMs on visual reasoning; and (iii) the public release of the dataset, evaluation code, and annotations to support future research.

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 have driven progress, they primarily involve single images with text-based answers, making models susceptible to linguistic priors rather than robust visual grounding (Goyal et al., 2017; Agrawal et al., 2018; Liu et al., 2024). More recent benchmarks extend VQA to pattern recognition (Xu et al., 2025), chart understanding (Masry et al., 2022), and mathematical reasoning (Zhang et al., 2024). Several works also explore multi-image or sequential reasoning (Imam et al., 2025; Vivoli et al., 2025), though these often emphasize temporal continuity or abstract transformations rather than fine-grained reasoning over discrete static image 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). Other more recent benchmarks target specific narrative phenomena, including humor understanding (Hu et al., 2024), scene-level classification (Paval et al., 2025), and abstract symbolic reasoning(Wang et al., 2025b). 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

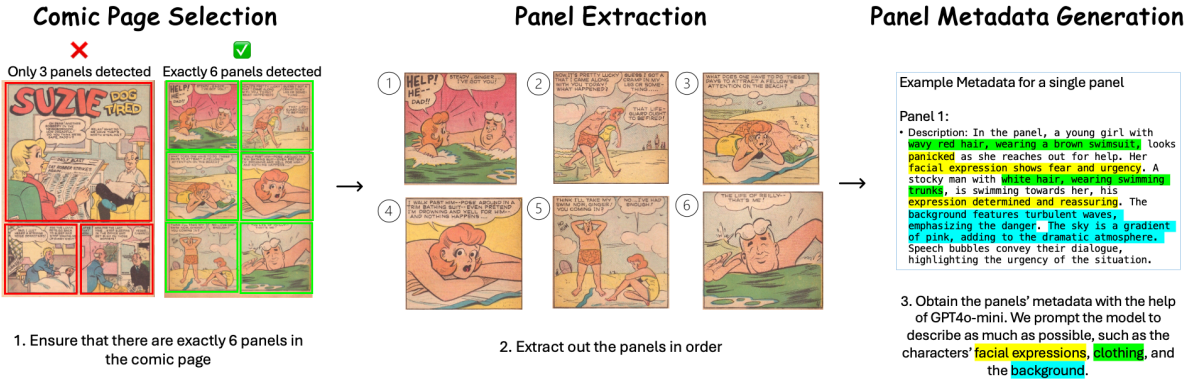

Figure 2: ComicVQA dataset construction consists of three steps: 1. Filtering down the comic pages to the ones with exactly six panels, 2. Extraction of panels from the finalized comic pages, and 3. Generate metadata for each panel for analysis. The prompt used to genereate the metadata is shown in Appendix A.

multiple-choice question with four options.

Task 1: Missing Panel Prediction. Given an incomplete comic strip with one panel removed, the model must select the panel that correctly fills the missing position. This task evaluates visual grounding and discrimination. To minimize reliance on textual cues, all dialogue and textual content are removed from panels, ensuring that models rely on perceptual signals such as character appearance, artistic style, and scene continuity. Candidate panels are selected to be semantically similar to the ground-truth panel, encouraging models to reason visually rather than using superficial semantic overlap. An example is shown in Figure 3.

Task 2: Panel Sorting. Given four shuffled comic panels, the model must identify the correct narrative order from four candidate permutations. This task evaluates compositional temporal reasoning, requiring models to integrate visual and textual information across panels to infer temporal progression and causal relationships within a coherent narrative. Questions contain predefined permutations, which allow for controlled comparison while maintaining enough variability to test global sequence understanding.

For both tasks, the model input consists of a task-specific prompt and a composed question image. Models produce free-form textual outputs, which are then mapped to the corresponding multiple-choice options for evaluation.

3.2 Dataset

3.2.1 Dataset Construction

The ComicVQA dataset is constructed from a subset of the COMICS dataset (Iyyer et al., 2017), comprising comic panels sourced from the public-domain Digital Comics Museum (DCM)[1], a repository of Golden Age comic books no longer under copyright. 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 consistency in input and question design, we select pages containing exactly six detected panels. Although individual questions involve only four or five panels, retaining six preserves narrative continuity and reduces the risk of disrupting the underlying story. All selected panels are recomposed into composite question images and resized to a maximum resolution of 1024×1024, to ensure uniform inputs.

In addition to visual data, we include metadata for each panel: a 100-word description generated with GPT-4o-mini (Hurst et al., 2024). These descriptions are used to select textually similar panels when constructing candidate options for the Missing Panel Prediction task, ensuring the task emphasizes visual reasoning. The full dataset construction pipeline is illustrated in Figure 2.

3.2.2 Question Design

We present all panels within a single composite image rather than 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

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

Figure 3: Example of a Missing Panel Prediction question. The correct option (2) can be identified through visual cues like the man's appearance and artistic style.

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 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 F.1.

3.2.3 Dataset Manipulation for Missing Panel Prediction Task

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. Additionally, all textual content was detected using EasyOCR (JaidedAI, 2024) and subsequently masked from the panels. This ensures that the models are forced to rely solely on visual information when discriminating between candidate panels. Figure 3 shows the final product after manipulation.

3.3 Dataset Statistics

The ComicVQA dataset consists of 6,157 multiple-choice questions constructed from comic strips. It includes two tasks: Missing Panel Prediction and Panel Sorting. Missing Panel Prediction contains 1,146 training, 511 validation, and 1,000 test questions, while Panel Sorting comprises 2,000 training, 500 validation, and 1,000 test questions.

Each question is associated with a composed image of comic panels, depending on task structure. In Missing Panel Prediction, one panel is masked and the masked position is evenly distributed across all panel positions. For both Missing Panel Prediction and Panel Sorting, the correct answer is randomly assigned to one of the four candidate options, ensuring balanced answer positions and a random baseline of 0.25.

All textual content in the comic panels, including narration and dialogue, is provided in English.

3.4 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. The prompts used for each task are shown in Appendix 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), InternVL-3.5 (Wang et al., 2025a), GPT4.1 (OpenAI, 2025), Claude-Sonnet-4.5 (Anthropic, 2025), and Gemini-Flash-2.5 (Comanici et al., 2025). All models are instruction tuned.

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 B.

Human Evaluation We conduct human evaluation to assess task clarity and establish an upper bound on performance. For each task, we uniformly sample 100 questions from the test set. For

Category	Model (#Params)	Missing Panel (%)	Panel Sorting (%)
Human	5 annotators	83.4	88.0
Baseline	Random	25.0	25.0
Open-Source	LLaVA-v1.6-Mistral-7B (7.57B)	23.9	25.1
	Qwen2.5-VL-7B-Instruct (8.29B)	30.2	26.8
	LLaMA-3.2-Vision-Instruct (10.7B)	25.2	26.6
	LLaVA-v1.6-Vicuna-13B (13.4B)	25.0	25.4
	Gemma-3-27b-it (27.4B)	35.8	26.9
	InternVL-3.5 (38B)	47.7	26.3
Proprietary	GPT-4.1	58.5	36.1
	GPT-5.1	62.6	33.7
	Claude-Sonnet-4.5	47.4	35.9
	Gemini-Flash-2.5	58.5	46.4

Table 2: Accuracy (%) of instruction-tuned models and human annotators on ComicVQA. Missing Panel Prediction evaluates fine-grained visual grounding, while Panel Sorting assesses compositional temporal reasoning.

Missing Panel Prediction, we additionally ensure balanced sampling by selecting 20 questions from each missing-panel position. Five annotators participated in the study, and the same annotators were used across both tasks.

For Missing Panel Prediction, annotators were asked to select the most appropriate comic panel from 4 options to fill in the missing panel in a comic strip. As for Panel Sorting, annotators were asked to select the correct ordering of four panels. Annotators also reported their confidence on a three-point scale (guessing, moderate confidence, very confident) for each of the question answered. Additional details on annotator instructions and study setup are provided in Appendix D.

4.2 Main Results

Table 2 summarizes performance on the two ComicVQA tasks across a range of open-source and proprietary MLLMs, together with human performance and a random baseline. Human annotators perform strongly on both Missing Panel Prediction (83.4%) and Panel Sorting (88.0%), with substantial inter-annotator agreement (Fleiss' κ = 0.712 and 0.757, respectively) and high confidence scores. This confirms that both tasks are well-defined and reliably solvable, while leaving a large margin for model improvement.

Missing Panel Prediction. Models achieve meaningful gains over random guessing, indicating that they can exploit visual information to perform panel discrimination. However, performance varies widely across models. Open-source MLLMs generally remain below 50% accuracy, while the strongest proprietary models reach only around 63%, still around 20 percentage points behind human performance. This persistent gap suggests that, even when the narrative structure is partially constrained, current models struggle with precise visual grounding and stylistic consistency across comic panels.

Panel Sorting. Panel Sorting is substantially more challenging. Most open-source models perform near chance, and even large proprietary models exhibit only modest improvements. Gemini-Flash-2.5 achieves the best performance at 46.4%, yet remains far below human accuracy. Unlike Missing Panel Prediction, where models benefit from localized visual cues, Panel Sorting requires integrating information across multiple panels to infer a globally coherent narrative, demonstrating a bigger challenge to solve this task.

Overall, these results suggest that while current MLLMs can leverage visual information for local discrimination in Missing Panel Prediction, they struggle to robustly infer global narrative structure in Panel Sorting. In the following sections, we analyze these behaviors in greater detail, focusing on model sensitivity to different types of errors.

5 Analysis of Model Performance on Panel Sorting

Overall performance on the Panel Sorting task is close to random for most evaluated models, with the exception of Gemini-Flash-2.5, which achieves the highest accuracy at 46.4%. In contrast, human annotators reach 88.0% accuracy, indicating a substantial gap between current multimodal mod-

Comparison	Perturbation Type	Accuracy (%)
12 vs **21**	Adjacent (1 ↔ 2)	64.7
23 vs **32**	Adjacent (2 ↔ 3)	64.5
34 vs **43**	Adjacent (3 ↔ 4)	65.7
	Average:	*65.0*

Table 3: Diagnostic accuracy of 2 panels across different perturbation settings. The model achieves an average accuracy of 65.0%, consistently above the 50% random baseline, indicating basic local temporal reasoning capability.

Comparison	Perturbation Type	Accuracy (%)
1234 vs **21**34	Adjacent (1 ↔ 2)	63.7
1**23**4 vs 1**32**4	Adjacent (2 ↔ 3)	69.0
12**34** vs 12**43**	Adjacent (3 ↔ 4)	69.3
	Average:	*67.3*
1234 vs **4**23**1**	Global (1 ↔ 4)	74.4

Table 4: Diagnostic accuracy of 4-panel binary comparisons. The model detects global endpoint swaps (1↔4: 74.4%) more reliably than adjacent swaps (67.3% avg), indicating coarse-grained temporal discrimination being better than fine-grained temporal discrimination.

els and human-level multimodal understanding in comics. To better understand the sources of model failure and identify challenges for future work, we conduct analysis focusing on Gemini-Flash-2.5, the strongest-performing model on this task.

5.1 Controlled Ordering Ablations

To better understand why models struggle with full four-panel sorting, we conduct a series of controlled ordering ablations that progressively simplify the task while preserving its core temporal reasoning requirements.

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.

We first analyze simplified two-panel ordering tasks, where the model must choose the correct order between two panels. From Table 3, across swaps of panels (1↔2), (2↔3), and (3↔4), the model achieves accuracies of 64.7%, 64.5%, and 65.7%, respectively, all well above the 50% random guessing score. This indicates that the model is able to capture local coherence when the comparison space is limited. An example of a two-panel comparison question is provided in Appendix G.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 also reduces the task to a two-option comparison, similar to the two-panel case, but with additional surrounding context. From Table 4, accuracy for swapping the first two, middle two, and last two panels is 63.7%, 69.0%, and 69.3%, respectively, with a peak of 74.4% for the global swap of the first and last panels (1↔4), suggesting that the model more easily distinguishes large narrative disruptions than subtle differences between adjacent panels. An example of a four-panel comparison question is provided in Appendix G.2.

Key patterns and implications. Two consistent patterns emerge from the binary discrimination experiments for Gemini-Flash-2.5. First, the model exhibits minimal benefit from additional context, with accuracy improving by only 2.3 percentage points from two-panel (65.0% average) to four-panel (67.3% average) binary tasks with adjacent swaps, indicating limited use of surrounding panels. Second, the model performs better on coarse-grained global swaps (74.4%) than on adjacent swaps (67.3%), showing reliance on coarse temporal heuristics and difficulty with fine-grained visual reasoning. Therefore, low performance on full panel sorting could stem from subtle adjacent transitions are often misordered, and insufficient context integration prevents the model from correctly sequencing all panels in the set.

5.2 Error Taxonomy for Panel Sorting failures

The binary controlled ordering ablations for Gemini-Flash-2.5 revealed difficulty with fine-grained local transitions, limited use of context, and reliance on coarse temporal heuristics. While these patterns indicate where the model struggles in full four-panel sorting, they do not reveal the specific types of reasoning challenges involved. To investigate this further, we categorize errors based on different aspects of panel understanding, analyzing mistakes in causal reasoning, character and object state tracking, scene transitions, and dialogue. We also examine how each error type relates to the temporal span of the panels.

Error Category	Description
Causal Reasoning	Failure to infer cause–effect relationships across panels.
Character State Tracking	Incorrect tracking of character identity, emotion, or pose across panels.
Object State Tracking	Failure to track object presence or movement across panels.
Scene Transition	Misinterpretation of scene changes, viewpoint shifts, or temporal context between panels.
Dialogue-Related	Misinterpretation on dialogue or narration, resulting in incoherent flow.
Ambiguity	Cases where multiple answer options are plausibly correct, even for humans.

Table 5: Failure taxonomy for the Panel Sorting task. Each incorrect model prediction was assigned exactly one dominant error category by two human annotators. Categories were derived by inspecting common model errors and reflect the most frequent failure patterns.

5.2.1 Annotation Setup

We also focus on Gemini-Flash-2.5, the strongest-performing model on Panel Sorting, to ensure our analysis highlights intrinsic model limitations rather than trivial failures. From the test set, we randomly sample 100 questions that the model answered incorrectly. Two human annotators who are selected based on their strong performance in the earlier human evaluation on the Panel Sorting Task, independently examined each incorrect prediction and assigned exactly one error category corresponding to the dominant failure mode. Additionally, each annotator labeled the temporal span of the error, indicating whether the failure involved local (error occurs within 2 adjacent panels), short-range (error spans 3 consecutive panels), or global (error involves the entire 4-panel sequence). A third annotator resolved any disagreements to produce consensus labels.

Error categories were derived inductively by inspecting common patterns in model predictions and are intended to capture the most frequent and salient failure modes. The different failure modes are shown in Table 5.

5.2.2 Error Distribution

Figure 4 shows the distribution across six error categories. Overall, from the bar plot, causal reasoning errors are the most common form of error (34%), followed by character state tracking (24%), object state tracking (14%), scene transitions (12%), and dialogue-related errors (11%). Notably, ambiguity errors are rare (5%), confirming that our ground truth orderings are clear and deterministic.

5.2.3 Error distribution by Temporal Span.

To further understand how model failures relate to the temporal extent of the error, we categorize each incorrectly answered test question by the span of panels involved: local, short-range, or global. These are then represented by the stacks on each of the bars in Figure 4.

Causal Reasoning Errors. Causal reasoning errors are concentrated in local spans (24 of the 34 questions), reflecting the difficulty models face in resolving adjacent-panel swaps where fine-grained temporal and causal reasoning is required, as compared to the more obvious changes across all four panels globally (4 of the 34 questions).

Character and Object State Tracking Errors. In contrast, entity state tracking errors occur predominantly across global spans (20 of the 24 questions under character state tracking errors, and 10 of the 14 questions under object state tracking errors). These failures typically stem from having to monitor character pose, identity, or interaction over the entire sequence, and the presence or movement of objects over the entire sequence. Tracking of the change in the entities' states usually requires many contextual clues from previous panels to ensure the next state of the entity is correct. This pattern is similar to our ablation finding of minimal context benefit, whereby the model cannot effectively integrate surrounding context fail to track entities across full sequences, leading to global-span errors.

Scene transition errors. These errors are distributed across local (8 of 12 questions) and short-range (4 of 12 questions) spans, highlighting challenges in maintaining visual continuity when environments change subtly between panels, as compared to global where the changes are much bigger. These errors often occur when small positional or scene changes compound across consecutive panels, requiring attention to multiple temporal steps. This phenomenon is similar to the ablations, whereby the model performs much better on obvious global disruptions, where the transitions are huge, as compared to local adjacent disruptions.

Dialogue-related errors. Dialogue-related errors appear almost exclusively in global spans (9 of 11 questions). While it is easy to ensure the flow

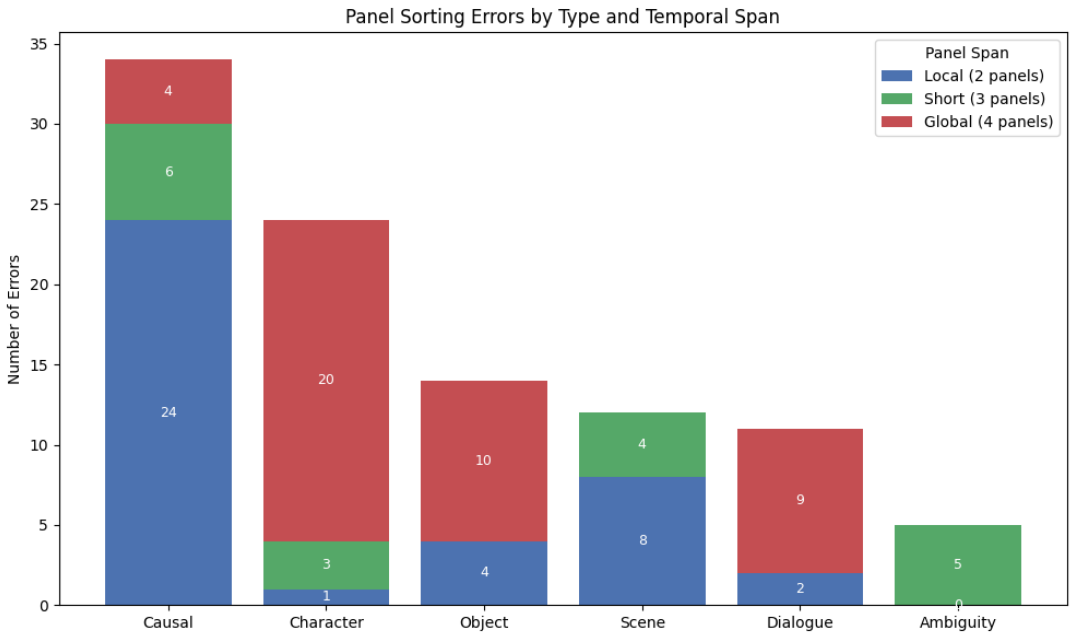

Figure 4: Stacked bar chart showing the distribution of error categories for Gemini-Flash-2.5 on the Panel Sorting task, stratified by the span of panels involved in each error. Bars show the percentage of errors attributed to each failure mode, with stacks indicating whether the error arises from local, short-range, or global reasoning.

of dialogue across adjacent panels, having to sort out 4 panels of dialogue is more challenging.

Overall, the combination of ablations and error-span analysis demonstrates that failure modes are structured, span-dependent, and non-trivial, directly supporting ComicVQA's role as a diagnostic benchmark for multimodal narrative reasoning. Gemini-Flash-2.5 struggle most with local, fine-grained reasoning while handling large, obvious global disruptions relatively well, providing directions for future improvements in causal reasoning, state tracking, and multi-panel integration.

5.3 Summary of Analysis

The model performs best on sequences with obvious global disruptions and slightly lower on subtle local reorderings, explaining the performance gap relative to human annotators. Performance on the Missing Panel Prediction task (58.5%) shows that the model can discriminate visual differences well above random chance (25%), but it remains far from human-level (88%) and has limitations. Overall, these findings highlight that the main challenge lies in reasoning over subtle local transitions, suggesting that future models may benefit from mechanisms that explicitly track evolving visual states and causal dependencies across panels.

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 with the absence of text in the comic panels, 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 in comics. 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, building manuals), and leveraging advanced multi-modal pretraining to improve complex visual and contextual reasoning.

Limitations

ComicVQA currently includes only English-language Western comics, primarily because publicly available comic datasets in other languages

or regions are limited or restricted by copyright. While this choice allows for open sharing and reproducibility, it also means that the dataset may not capture the full diversity of visual narrative styles across cultures. An example of a comics dataset that are restricted by copyright is the Manga109 (Aizawa et al., 2020) dataset. This dataset is only allowed for academic purposes only. We would like our dataset to be used as part of training of future models to further improve on the model's visual reasoning in the area of comics and beyond. Future work could extend the benchmark to other languages, regions, and artistic styles when public data becomes available.

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.

The segmentation of the comic panels are not perfect. Comic artists do not need to explicitly follow the usual square-like panels of the same size. Therefore, the segmentation algorithm needs to account for the various shapes like circles and rectangles, and even odd shapes like stars and triangles. As a result, the cut of the panels are not perfect, meaning there might still be some information (textual and visual) loss or noise added for each panel. However, despite this limitation of the panel segmentation algorithm, a high human baseline score shows that the tasks are still solvable to a large extent.

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.

Anthropic. 2025. System card: Claude sonnet 4.5. Accessed: 2026-01-03.

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.

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

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*.

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. *Advances in Neural Information Processing Systems*, 37:47166–47188.

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.

JaidedAI. 2024. Easyocr.

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.

Sandro Paval, Pascal Meißner, and Ivan P. Yamshchikov. 2025. ComicScene154: A scene dataset for comic analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31562–31568, Suzhou, China. Association for Computational Linguistics.

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.

Yuriel Ryan, Rui Yang Tan, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2025. Humor in pixels: Benchmarking large multimodal models understanding of online comics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14024–14050, Suzhou, China. Association for Computational Linguistics.

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*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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

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 Prompts used in this paper

Prompt for Generation of Panel Metadata
Describe the panel's visual content thoroughly, including the characters' poses, facial expressions, actions, clothing, background, and any visual storytelling cues. Maximum 100 words.

Prompt for Missing Panel Prediction
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:

Prompt for 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:

B Model Hyperparameters and Evaluation Protocol

B.1 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.

For proprietary models, they were evaluated using OpenRouter[3] API, and to ensure a fair comparison with open-sourced models, `reasoning_effort` was set to none and `thinking_budget` was set 0 zero to ensure that there is no thinking/reasoning tokens.

B.2 Experiment Settings

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

B.3 Time Taken

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.4 Accuracy Evaluation

For all tasks in ComicVQA, each question has a single correct answer selected from a fixed set of candidates.

- Missing-panel task: Four candidate panels, one correct panel per question.
- Panel-ordering task: Four candidate permutations of the panels, one correct permutation per question.

Accuracy is computed per question: a prediction is counted as correct if the model's selected option matches the ground-truth answer, and incorrect otherwise. Random guessing yields 25% accuracy for both tasks. All reported results in Tables 6 and 2 follow this criterion.

[3] https://openrouter.ai

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[4].

Digital Comics Museum is a publicly available data source for Golden Age comics.[5]

[4] https://github.com/miyyer/comics/blob/master/LICENSE

[5] https://digitalcomicmuseum.com/forum/index.php

D Human Evaluation

D.1 Basic Information on Human Annotators

We recruited five annotators from our professional network. Each was contacted individually via email or messaging and provided with information about the expected time commitment (approximately 1–1.5 hours) and compensation (approximately 25USD upon full completion). All annotators were native or fluent English speakers with no prior exposure to the ComicVQA dataset. Participation was voluntary, and all annotators agreed to complete the tasks under these conditions, indicating that the compensation was acceptable for the effort required.

D.2 Human Evaluation Protocol

Before starting on human evaluation, the annotators are to consent to the information provided to them regarding compensation, rights, and data usage.

Consent to participate in this study
STUDY PURPOSE: We are evaluating human performance on visual reasoning tasks involving comic book panels. Your responses will help establish a baseline for comparing human and machine performance on comic understanding tasks.
WHAT YOU'LL DO:
- Complete 200 multiple-choice questions
- Two types of tasks: (1) Identify missing panels, (2) Order shuffled panels (that will be in another form) - Estimated time: 20-30 mins for task 1, 30-45mins for task 2
- You can take breaks and resume later
YOUR RIGHTS:
- Participation is voluntary
- You may withdraw at any time without penalty (please let me know if you decide to withdraw)
- Your responses will be anonymized
- No personally identifiable information will be collected beyond what's required for payment
COMPENSATION:
- Full completion: 25USD
- If you take much longer than expected (for eg, 2 hours) please let me know and I will further compensate you.
DATA USAGE:
- Your responses will be used for academic research only
- Results may be published in academic venues (anonymized)
- Data will be stored securely and used only by authorized researchers
- No responses will be linked to your identity in any publication
If you agree to the above conditions, please click 'Yes' and proceed to work on the questions. If you choose not to proceed with the evaluation, please click 'No' and submit the form.

Questions were presented via Google Forms showing the exact same question image that is being fed to the models during inference time. Annotators were tasked to select option 1,2,3 or 4, with the below instruction:

Instructions for Missing Panel Prediction
TASK 1: MISSING PANEL PREDICTION
You will see 4 comic panels arranged in sequence, with one panel marked as [?] in a grey box. Your task is to identify which of 4 options correctly fills the missing position.
IMPORTANT NOTES:
- Text has been REMOVED from all panels in this task intentionally
- Select the option that best completes the sequence
- There is only ONE correct answer
- Must choose your confidence level
Please answer based on what you think is correct. Don't overthink it!

Instructions for Panel Sorting
TASK 2: PANEL SORTING
You will see 4 shuffled comic panels. Your task is to determine the correct reading order from 4 possible options.
IMPORTANT NOTES:
- Panels include original text/dialogue
- Think about the natural flow of the story
- Western comics typically read left-to-right
- There is only ONE correct answer
- MUST give me a confidence rating
Please answer based on what you think is correct. Don't overthink it!

Additionally, they were asked on their confidence in answering the question. Figure 5 shows a screenshot on how each question looks like on the form.

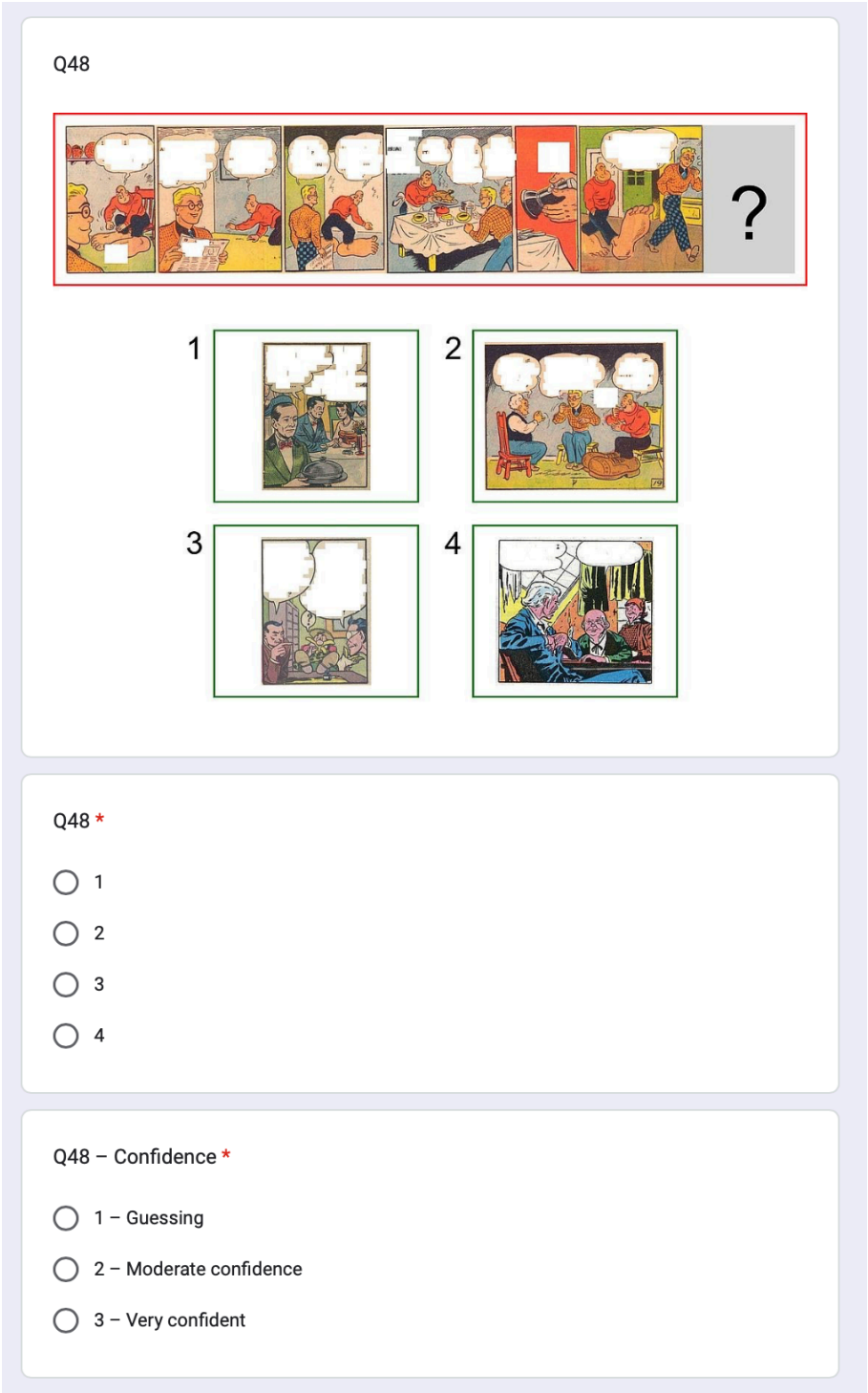

Figure 5: Example question.

D.3 Human Annotation Protocol

Human annotation protocol is for the three annotators that are assigned to work on determining the failure type for each of the 100 questions sampled.

Similar to human evaluation protocol, before starting on the human annotation, participants are to content to the information provided to them regarding compensation, rights, and data usage. Compensation for this was also approximately 25USD for full completion.

Instructions for Annotation of Failure Types
TASK: ANNOTATE THE FAILURE TYPE
You will see a Panel Sorting Task question, and below the image will indicate what the option the model chose, and what the correct answer actually is.
You are to answer 2 main questions, main error type, and the extent of the error.
Here are the list of plausible error types (these are the common error types):
1. Casual Reasoning: This happens when models fail to infer cause-effect relationships across panels. For example, panel shows broken glass before another panel showing the glass about to land on the floor.
2. Character State Tracking: This happens when the model incorrectly tracks the character's identity, emotion, or pose across panels.
3. Object State Tracking: This happens when the model fails to track object presence or movement across panels.
4. Scene Transition: This happens when there are scene changes or viewpoint shifts, but the model did not catch the changes based on the surrounding panels.
5. Dialogue Related: This happens when dialogue flow of the option that the model selected is obviously not right.
6. Ambiguity: Sometimes, the option that the model selected is also plausible. You can mark it as ambiguity if the option the model selected makes sense to you.
After choosing the error, please share with us how much of this error is present in the panels.
Guidelines on the options given:
1. Local: The error occurs within 2 adjacent panels
2. Short-range: Error spans 3 adjacent panels
3. Global: Error spams the whole entire 4 panel sequence
Please answer based on what you think is correct.

Figure 6 shows how the question look like on the form.

Q14 – Failure Annotation
Correct ordering: Option 2

Model's Prediction: Option 4

Please label the following based on the model's failure.

Q14 – Dominant Failure Type *

- ◯ Causal Reasoning
- ◯ Character State Tracking
- ◯ Object State Tracking
- ◯ Scene Transition
- ◯ Dialogue-Related
- ◯ Ambiguous choices

Q14 – Error Span
Please label the extent of the error

Q14 – Error Span *

- ◯ Local
- ◯ Short-range
- ◯ Global

Figure 6: Example question.

Approach	Accuracy (%)	Gap to Random (%)
Vision-only (CLIP ViT-L/14)	35.6	+10.6
Vision-only (DINOv2 ViT-L/14)	40.2	+15.2
Vision-only (SigLIP-Large)	46.1	+21.1
Random baseline	25.0	–
Best MLLM (GPT-5.1)	**62.6**	**+37.6**

Table 6: Performance of vision-only baselines on the missing-panel task. Vision-only models outperform random guessing but fall substantially short of the best multimodal model, indicating that visual similarity alone is insufficient for reliable narrative reasoning.

E Can vision encoders naively solve the Missing Panel Prediction Task?

A natural question is whether the missing-panel task can be solved using visual similarity alone, without any language modeling component. In particular, one might hypothesize that a vision transformer embedding all panels could select the correct answer via cosine similarity, based on cues such as artistic style or character appearance.

To test this hypothesis, we evaluate three strong vision-only encoders: CLIP ViT-L/14, DINOv2 ViT-L/14, and SigLIP-Large. We use cosine similarity between the query panels and candidate responses to obtain the answer. A random baseline (25%) is included for reference.

As shown in Table 6, vision-only models achieve accuracies between 35.6% and 45.9%, substantially above random chance but far below the best multimodal model (62.6%).

These results indicate that vision-only models can exploit visual features, but relying on visual discrimination alone is insufficient. Weak MLLMs (23–30%, Table 2) fail to leverage visual cues effectively despite having access to them. Stronger open-source MLLMs, such as Gemma-3-27b-it and InternVL-3.5, outperform the vision-only encoders, suggesting that they integrate visual discrimination with higher-level reasoning over the structure and relationships of panels. Proprietary models, as the strongest performers, successfully combine both visual cues and visual reasoning, achieving the highest accuracy on this task.

F Example Questions from Panel Sorting Task

F.1 Panel Sorting Example Question

Figure 7: 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.

G Analysis Example Questions

G.1 Panel Sorting 2 Panel Example Question

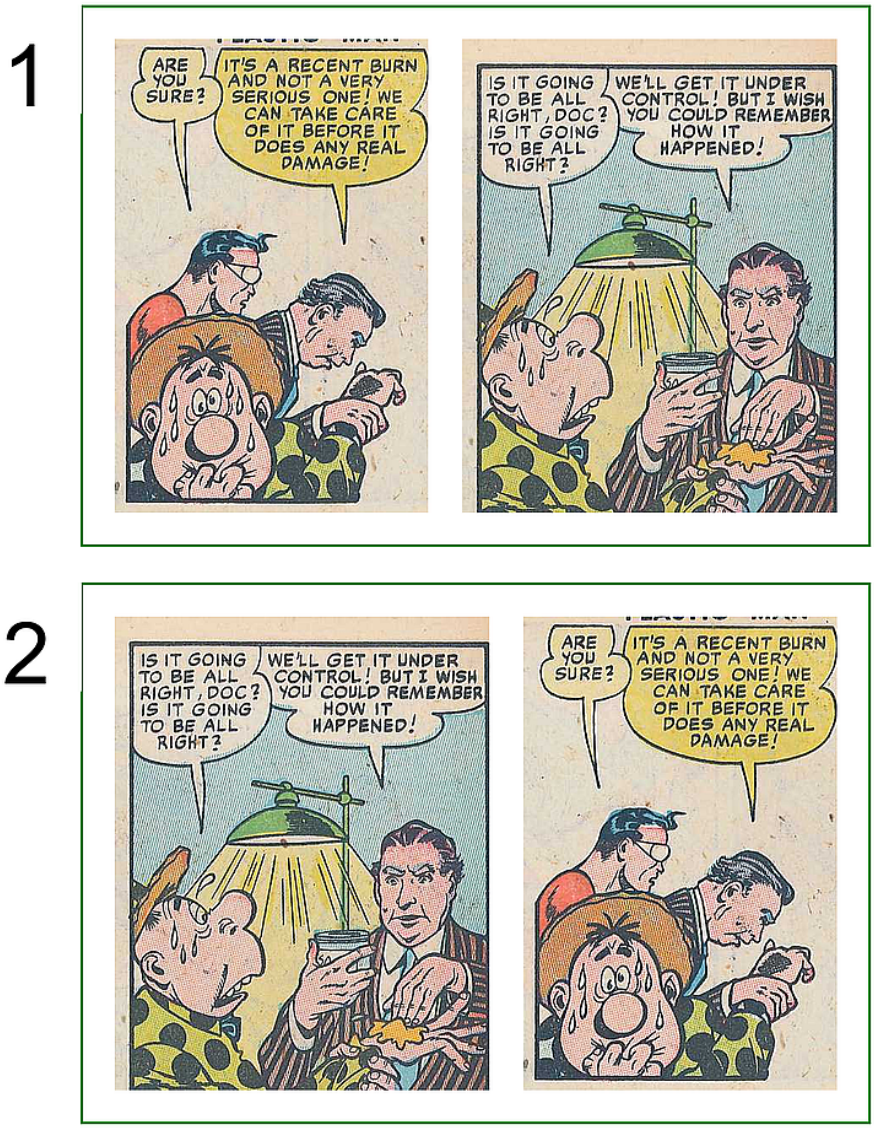

Figure 8: For analysis, we cut down the number of panels to see any changes in the performance of the model with a smaller number of candidate options. Since 2 panels only have 2 permutations, theres only 2 options.

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

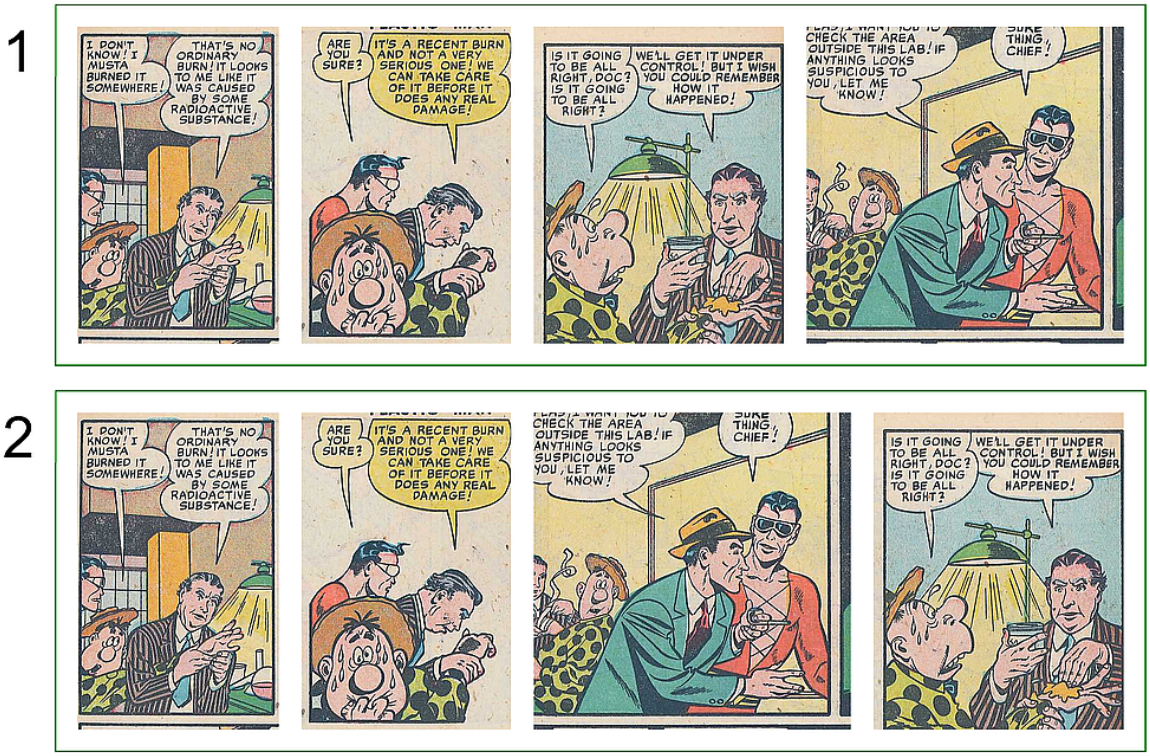

Figure 9: Simlar to the 2 panel analysis, we keep 2 panels in the correct order, and shuffled the remaining 2.