

---

# MangaVQA and MangaLMM: A Benchmark and Specialized Model for Multimodal Manga Understanding – Supplementary Material –

---

1 In this supplementary material, we provide additional details including (A) OCR evaluation in comics,  
2 (B) synthetic VQA examples, (C) setup details, and (D) additional results.

## 3 A OCR Evaluation in Comics

4 As described in §2, the evaluation of OCR has often been underexplored. Recent works such as  
5 Magi [1] and CoMix [2] focus on transcription generation, which inherently includes OCR as a core  
6 component. CoMix, in particular, proposes a dedicated metric called the Hybrid Dialog Score for  
7 evaluating transcription tasks. However, this transcription-focused evaluation differs from direct OCR  
8 evaluation, which aims to assess whether the model accurately reads the text. First, transcription  
9 involves multiple subtasks beyond text detection and recognition, such as speaker identification,  
10 reading order prediction, and others. The quality of the final transcription output depends on the  
11 combined performance of these components, making it difficult to isolate and measure the accuracy  
12 of text recognition alone.

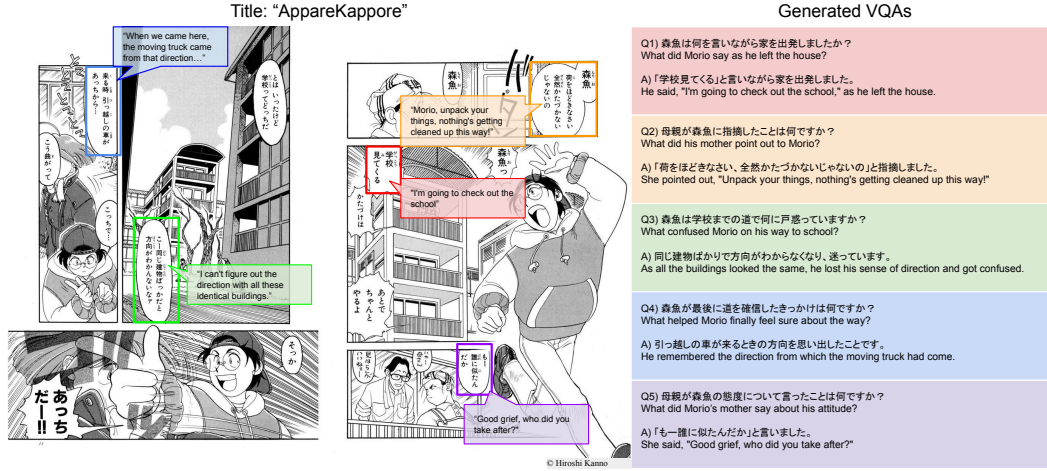
13 Second, transcription-based evaluations do not assess the positional accuracy of recognized text.  
14 Spatial information plays a crucial role in OCR, especially when the same text appears in multiple  
15 locations, as it helps identify which text instance is correct. For example, in Figure E(a), the word  
16 “わあー (waa-)” appears in four different locations, only one of which is correct. Without positional  
17 information, it becomes impossible to identify the correct instance. Moreover, spatial information  
18 is crucial for content understanding, as the interpretation of the same text can vary significantly  
19 depending on its location.

20 A proper evaluation of OCR in the manga domain allows us to better understand how well current  
21 LMMs can recognize text within manga. As described in the results section (§6.1), models such as  
22 GPT-4o exhibit near-zero OCR performance, yet are still able to answer VQA questions that rely  
23 on textual information. This result suggests that LMMs may be partially recognizing some text in  
24 the image. Our visualization of GPT-4o’s OCR output reveals that the detected text regions almost  
25 always appear in nonsensical locations, yet the model can still read certain parts of the text in the  
26 image. We provide a detailed analysis of this observation in §D.2.

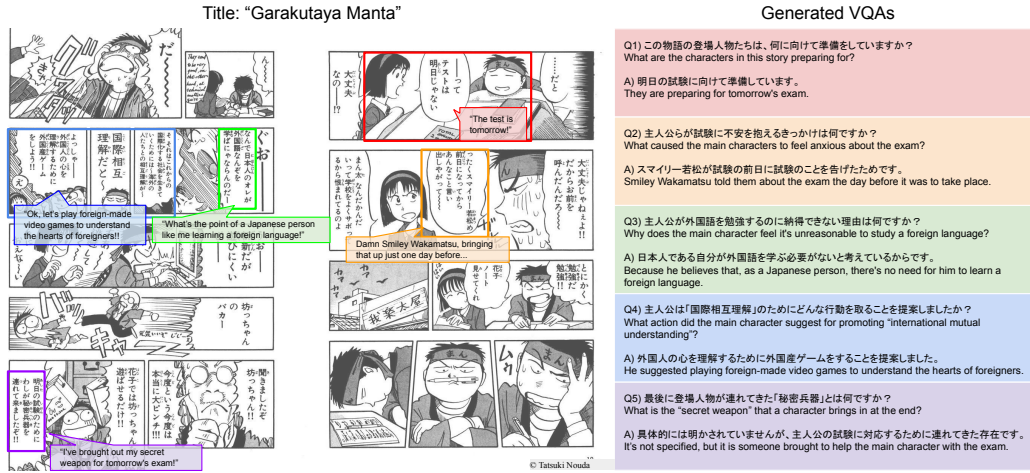
## 27 B Synthetic VQA Examples

28 For training our MangaLMM, we rely on synthetic VQAs generated by GPT-4o. In Figure A, we  
29 provide examples of these generated VQAs. As illustrated in the figure, GPT-4o is capable of  
30 producing accurate and diverse question–answer pairs.

31 We emphasize once again that providing GPT with text annotations is crucial for generating such  
32 high-quality VQAs. Without these annotations, GPT tends to produce unreliable outputs (e.g.,  
33 misspelled extractions and factually incorrect questions) which significantly limit the performance of  
34 the MangaLMM trained on such data, as discussed in §6.4.



(a) An example from the manga titled AppareKappore.



(b) An example from the manga titled GarakutayaManta.

Figure A: Examples of synthetic VQA generation results. The most relevant part of the image for each question-answer pair is highlighted and translated in the corresponding color.

35 **Human Validation.** To validate the reliability of the synthetic VQA data generated by GPT, we  
 36 conducted a manual evaluation. We randomly sampled 100 question-answer pairs and asked four  
 37 human evaluators to assign scores to each item on a three-level scale: 0 (incorrect), 0.5 (partially  
 38 correct), and 1 (correct). The average score is 0.78, suggesting that approximately 80% of the  
 39 synthetic VQAs are judged to be appropriate by humans.

## 40 C Setup Details

41 **Evaluation Metric.** We provide a detailed description of the normalized edit distance (NED, also  
 42 referred to as 1-NED), which was used as the evaluation metric in MangaOCR. NED scales the  
 43 standard edit distance to a range between 0 and 1, where higher values indicate better prediction. It is  
 44 computed as follows:

$$\text{NED} = 1 - \sum_{i=1}^N \frac{\text{ED}(\text{GT}_i, \text{Pred}_i)}{\text{MaxLen}(\text{GT}_i, \text{Pred}_i)} \quad (1)$$

Here,  $GT_i$  and  $Pred_i$  denote the  $i$ -th ground truth and the model’s prediction, respectively.  $ED(\cdot)$  calculates the edit distance between two strings, and  $MaxLen(\cdot)$  returns the longer of the two string lengths.  $N$  indicates the total number of text instances.

## C.1 Prompt

**Prompt for Synthetic VQA Generation.** For creating synthetic QA pairs for training, we provide GPT-4o with the prompt in Table A along with the corresponding image.

Table A: Prompt for the synthetic VQA generation.

Original Japanese
<p>与えられる画像と、そこに書かれている文字情報を用いて、          質問: [質問内容]          回答: [回答内容]          質問: [質問内容]          回答: [回答内容]          ...          の形式でVQA問題を5問作ってください。解釈が曖昧になる主観的な問題ではなく、書かれている事実に基づいて客観的に判断できる問題を作ってください。またOCRのような文字の読み取り問題にはせず、内容理解を問う問題を作ってください。          画像内の文字:          {OCR ANNOTATION HERE}</p>
Translated
<p>Using the given image and the textual information written in it, create 5 VQA questions in the following format:          Question: [Question content]          Answer: [Answer content]          Question: [Question content]          Answer: [Answer content]          ...          Avoid subjective questions that could lead to ambiguous interpretations, and instead create questions that can be objectively answered based on the facts presented in the image. Also, do not include OCR-style text recognition questions; instead, create questions that test understanding of the image content.          Text in the image:          {OCR ANNOTATION HERE}</p>

**Prompt for Training and Evaluation.** For training and inference, we use task-specific prompts. For the MangaOCR benchmark, we provide the prompt “Please perform OCR on this image and output the recognized Japanese text along with its position (grounding)” along with the input image. During training, the corresponding OCR annotations are included as supervision. When running OCR inference with GPT-4o, Gemini 2.5, and Phi-4, the outputs varied in format unless explicitly specified. Therefore, we use the prompt in Table B to align their outputs with the OCR format used in the training data of MangaOCR.

For the MangaVQA benchmark, we use the prompt “あなたは日本語の漫画に関する質問に答えるAIです。与えられた画像に基づいて質問に答えてください。(You are an AI that answers questions about Japanese manga. Please answer the given question based on the provided image.)” together with the input image and a question. The ground-truth answer is given only during training. For MangaVQA evaluation, the prompt in Table C is used for LLM-as-a-judge.

Table B: OCR inference prompt for GPT-4o, Gemini 2.5, and Phi-4.

<p>Please perform OCR on this image and output the recognized Japanese text along with its position (grounding).</p> <p>The output should be a JSON list. Each item in the list must follow the structure below:</p> <pre>\n{"bbox_2d": [x1, y1, x2, y2], "text_content": "..."}\n</pre> <p>The field <code>"bbox_2d"</code> must be a 2D bounding box that tightly encloses the text. Use the format <code>[x1, y1, x2, y2]</code>, where:</p> <ul style="list-style-type: none"> <li>- <code>'x1'</code>, <code>'y1'</code> are the coordinates of the top-left corner of the bounding box, and</li> <li>- <code>'x2'</code>, <code>'y2'</code> are the coordinates of the bottom-right corner.</li> </ul> <p>Here is an example of the desired format:</p> <pre>\n{"bbox_2d": [1490, 138, 1546, 201], "text_content": "春休みです-"}\n</pre> <p>Please follow this format strictly.</p>
---

Table C: Prompt for MangaVQA evaluation.

System message
<p>You are an evaluator. Your task is to rate how appropriate a model's response is to a question about a manga image. For each case, you will be given a question (based on a manga image), a human-written answer, and the model's response. The image is not shown, but the question and answer are based on it. Please evaluate as if the image were available.</p> <p>Please rate how well the model's response answers the question, considering the intended image context and the human answer as a reference, using a scale from 1 to 10:</p> <ol style="list-style-type: none"> <li>1 — Completely inappropriate or unrelated to the question or image context.</li> <li>2 — Mostly unrelated with major misunderstandings or incorrect information.</li> <li>3 — Slightly relevant, but largely incorrect or unhelpful.</li> <li>4 — Somewhat relevant, but contains significant errors or omissions.</li> <li>5 — Partially correct with noticeable inaccuracies, vagueness, or missing key points.</li> <li>6 — Generally okay, but missing core points or includes some incorrect interpretations.</li> <li>7 — Mostly correct and relevant, with only minor issues or small omissions.</li> <li>8 — Almost entirely accurate with only slight room for improvement.</li> <li>9 — Very appropriate, accurate, and well-aligned with the question and image context.</li> <li>10 — Perfectly appropriate, accurate, and fully answers the question as if the image were visible.</li> </ol> <p>Only return a single number (1–10). Do not include any explanations, justifications, or comments.</p>
User prompt
<p>Input:</p> <pre>"question": {question}, "human-written answer": {answer}, "model's response": {generated_answer},</pre> <p>Your score:</p>


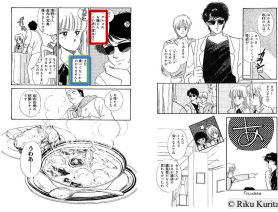
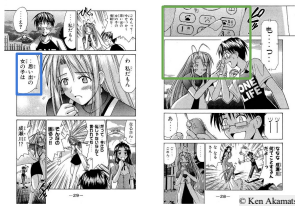
	Exact Extraction	Multimodal Understanding	Image Understanding
Question	2人を待ちくたびれているのは誰だと言っていますか？ Who is said to be tired of waiting for the two of them?	グルメのお嬢さんは店の採点について何と言っていますか？ What does the young gourmet lady say about the restaurant's rating?	右側のページにおいて、男の子が怒っているのはどうしてですか？ On the right-hand page, why is the boy angry?
Ground Truth	サタン様 Lord Satan	まだ食べていないからわからないと言っています。 She says she doesn't know yet because she hasn't eaten there.	成瀬川が突然勝手に電話を切ったから。 Because Narusegawa suddenly hung up the phone without warning.
			
Original Model	漫画の内容から、2人を待ちくたびれているのは「いかげん」と言っています。... Based on the manga's content, it says that "ikagen" is the one who is tired of waiting for the two... 3	グルメのお嬢さんは、「この店の採点は？」と尋ねています。 The young gourmet lady asks, "What's this restaurant's rating?" 2	右側のページでは、男の子が怒っている理由は、彼が何かを誤解している... On the right-hand page, the boy may be angry because he has misunderstood something... 4
Trained MangaLM	サタン様 Lord Satan 10	「まだ食べていないからわかりません！」と言っています。 She says, "I don't know yet—I haven't eaten here!" 10	女の子が「思い出の女の子」について言及したため。 Because the girl mentioned the "girl from his memories". 2

Figure B: **Category-wise analysis on MangaVQA.** The regions in the image relevant to the question or models' answer are highlighted with boxes in corresponding colors.

## D Additional Results

We provide additional analysis and experimental results on our benchmarks, MangaVQA and MangaOCR.

### D.1 More Analysis of MangaVQA

**Comparison with Human Evaluation.** To validate the reliability and consistency of the GPT-judge employed in the MangaVQA evaluation, we conducted a comparative analysis between its evaluation scores and those provided by human annotators. Specifically, we randomly sampled 100 items from the benchmark dataset and asked two human evaluators to assign scores to each item, following the same evaluation prompt used for the GPT-judge.

The results of this comparison are illustrated in Figure C. We observe a small absolute difference in average scores ( $\Delta = 0.22$ ). Additionally, there is a strong positive correlation between the scores assigned by the GPT-judge and the human average ( $r = 0.94$ ). These findings suggest that GPT-based evaluation can serve as a practical and consistent alternative to human judgment in our MangaVQA benchmark.

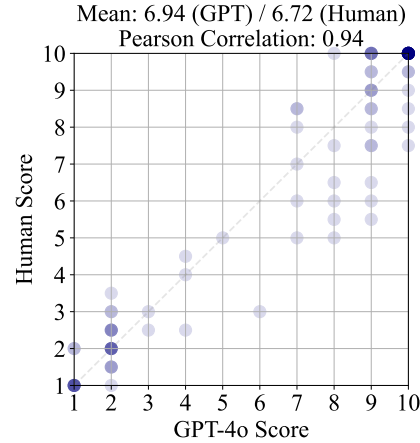


Figure C: **Comparison between GPT-Judge and Human Evaluation.** Darker points indicate a higher concentration of points.

**More Analysis of OCR Annotation when Generating VQA Data.** As described in §6.4, OCR annotation plays a key role in generating high-quality QA pairs with GPT-4o. Here, we provide a more detailed analysis of the effect of OCR annotation. OCR annotation consists of both bounding box positions and their text content. We compare the synthetic VQA data generated by GPT-4o using only the text content with those generated using both bounding box positions and text content. Table D presents the results. Interestingly, our experiments show that using only the text content is more effective than including both text and positional information. Although our current approach did not benefit from positional information,

Table D: Effect of OCR Annotation on VQA Generation.

OCR Annot.	LLM (/10.0)
None	5.44
Text	<b>6.57</b>
Text + Pos.	6.17

leveraging it remains a promising direction for future work. Therefore, in our experiments, we use synthetic VQA examples generated using only the OCR text content.

**Qualitative Analysis of MangaVQA.** Figure B presents category-wise examples on MangaVQA. For the categories on the left (Exact Extraction) and in the center (Multimodal Understanding), the base Qwen 2.5-VL model often fails to locate the correct region and consequently extracts the wrong words. However, these issues are significantly improved after training.

On the other hand, for the category on the right (Image Understanding), which does not rely on textual cues, MangaLMM tends to over-prioritise text extraction, leading to incorrect answers even after training.

## D.2 More Analysis of MangaOCR

We present a qualitative analysis of MangaOCR results from GPT-4o and MangaLMM. As described in §6, text segments that appear more than ten times are considered noise and excluded from the results. Therefore, such repeated segments do not appear in the visualizations.

**GPT-4o’s Results on MangaOCR.** Since previous studies have rarely conducted in-depth qualitative analysis of GPT-4o’s OCR results, it is difficult to assess the model’s actual performance on manga datasets. We address this gap by providing a detailed qualitative analysis of GPT-4o’s MangaOCR outputs. Figure D shows GPT-4o’s results on MangaOCR. These examples demonstrate the low zero-shot OCR performance of GPT-4o in the manga domain. The detected text regions almost always appear in incorrect or nonsensical locations, although the model can still read certain parts of the text within the image. Because the predicted text positions are inaccurate, the outputs are considered entirely incorrect under OCR evaluation criteria. While some predicted text fragments correspond to actual text in the image, there are many cases—such as in Figure D(b)—where most of the text is not recognized at all. Even when text is recognized, it is often incorrect. While GPT-4o fails to correctly detect and recognize most of the text, it can still recognize partial text content, which may allow GPT-4o to answer some text-based VQA questions.

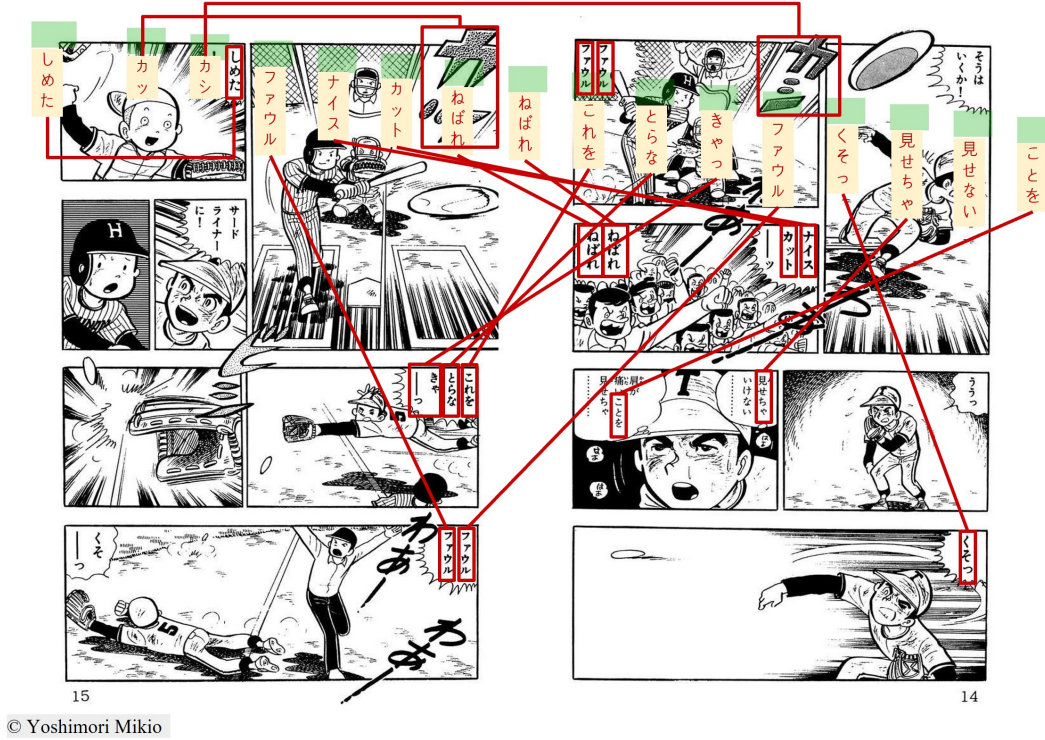
Interestingly, when performing OCR inference with GPT-4o, the model sometimes generates disclaimers such as: “The bounding box coordinates and text content are illustrative and may not perfectly match the actual image. For precise OCR and bounding box extraction, specialized OCR tools like Tesseract or Google Vision API should be used.” This suggests that GPT-4o itself acknowledges its limitations in precise OCR and recommends using dedicated OCR tools.

**MangaLMM’s Results on MangaOCR.** Figure E shows MangaLMM’s results on MangaOCR. As seen in the figure, most predictions appear correct, reflecting the model’s strong OCR capability across a wide range of text sizes, from large to small. The red regions indicate false negatives. Occasionally, even text that appears large and seemingly easy to detect is missed. According to our manual inspection, such cases are mostly onomatopoeia. This suggests that the model struggles more with onomatopoeic expressions, which are often written in non-standard fonts, sizes, or orientations, compared to regular text.

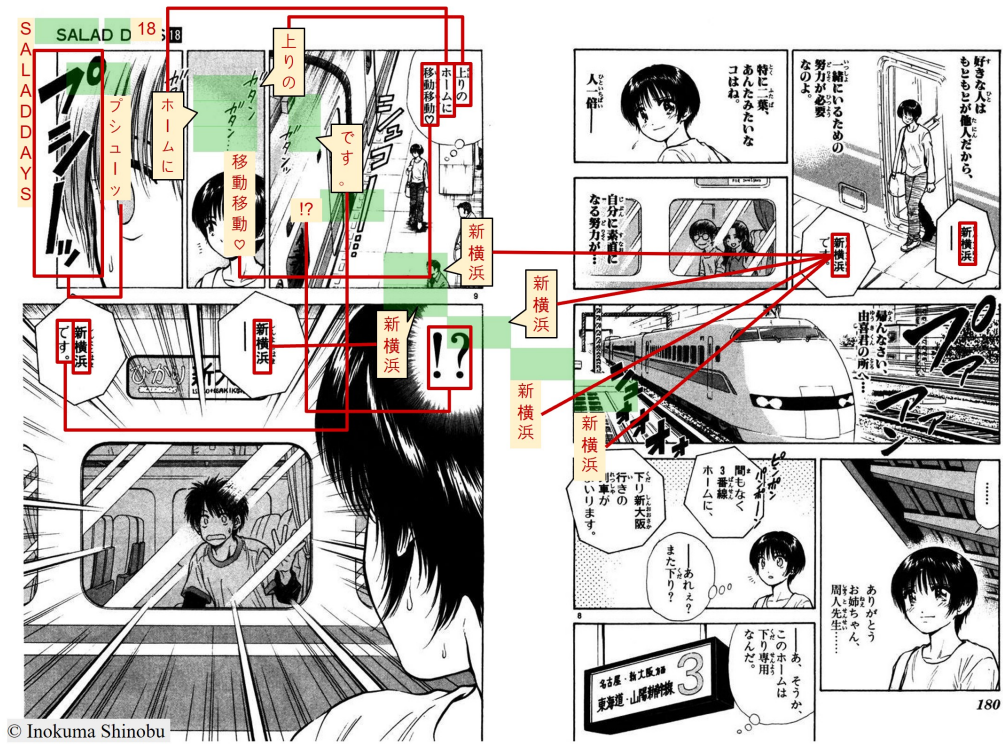
## References

- [1] Ragav Sachdeva and Andrew Zisserman. The manga whisperer: Automatically generating transcriptions for comics. In *CVPR*, 2024.
- [2] Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. Comix: A comprehensive benchmark for multi-task comic understanding. In *NeurIPS*, 2024.



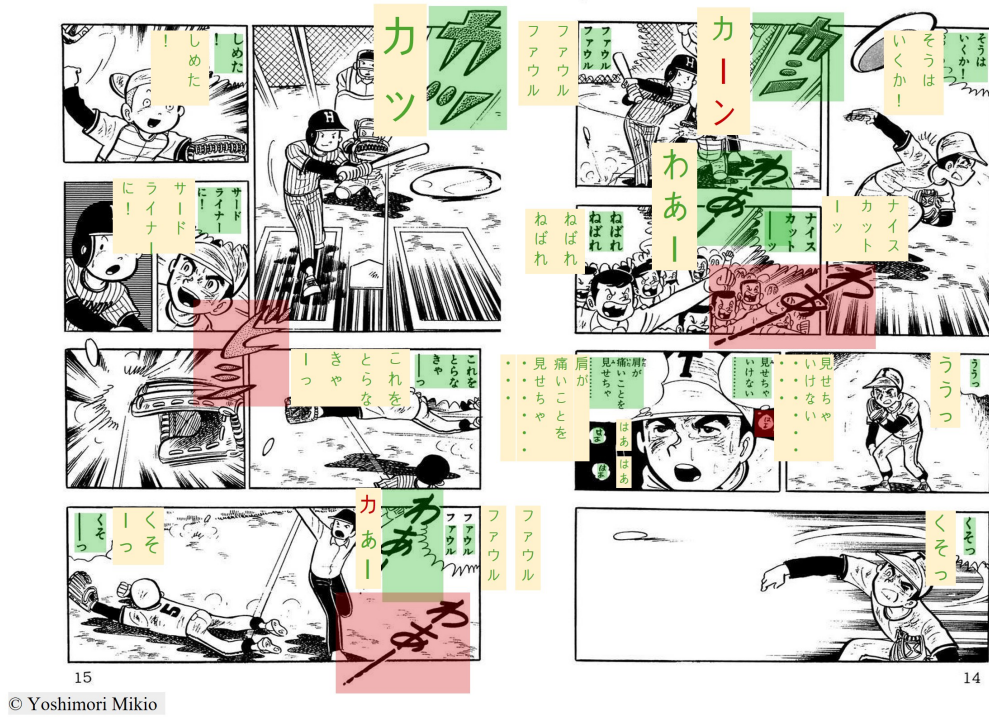


(a) An example from the manga titled ShimatteIkouze.

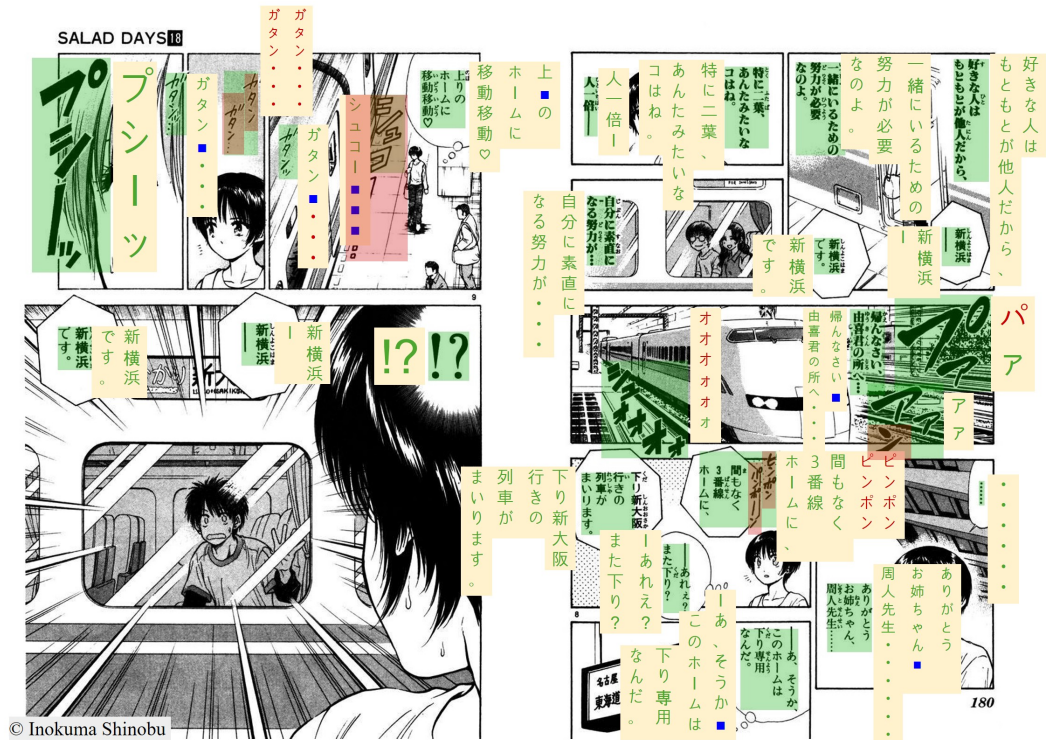


(b) An example from the manga titled SaladDays.

Figure D: GPT-4o's Results on MangaOCR. The green boxes indicate the detected text regions. The red text, shown near each green box, represents the predicted text fragment corresponding to that detected region. Each red bounding box is manually drawn to indicate where the predicted text fragment appears in the image. Red lines connect each predicted fragment to its corresponding detected position. These detected positions are almost always incorrect.



(a) An example from the manga titled Shimattelkouze.



(b) An example from the manga titled SaladDays.

Figure E: **MangaLMM’s Results on MangaOCR**. The green boxes indicate the detected text regions. The text shown near each green box is the predicted text for that detected region. The green text represents correctly predicted text, while the red text indicates incorrectly predicted text. Missing characters are marked with small blue squares. The red boxes show false negatives—text regions that should be detected but are missed. Most OCR results are correct.