

MangaLens: A Multi-Model Tool for Structured Manga Parsing

Jiannan Mao, Jin Tei, Zhi Qu, Hideki Tanaka, Masao Utiyama

National Institute of Information and Communications Technology, Kyoto, Japan

Abstract

Manga pages combine text, layout, characters, and narrative order, making them difficult to process as plain images or text alone. We present MangaLens, an integrated tool for structured manga parsing. MangaLens combines LLM-based reasoning with segmentation-based grounding to support manga parsing subtasks such as panel structuring, speaker linking, and related cross-page reasoning tasks. The system produces page-level and book-level JSON records and visualizes subtask outputs with aligned image evidence, enabling users to verify, correct, and reuse parsing artifacts. Evaluation results show that MangaLens produces reliable structured parsing outputs across major subtasks. The tool is released at <https://github.com/nict-astrec-att/MangaLens>.

1 Introduction

Manga is a widely consumed narrative medium that combines language, visual composition, and sequential storytelling (Cohn, 2014, 2016). As manga content is increasingly distributed through digital platforms and international markets (Grand View Research, 2025), there is a growing need for computational tools that support translation, annotation, and dataset construction. These applications require systems to process manga pages as structured multimodal documents. This is challenging because manga pages encode meaning through the interaction of words, images, spatial layout, and narrative order. Parsing a page therefore requires not only recognizing text, but also identifying where visual elements are located, how panels and speech bubbles should be read, who is speaking, and how characters and events are connected across pages (Sachdeva and Zisserman, 2024).

Existing datasets and benchmarks have supported important subtasks in manga and comic parsing, including Manga109 annotations for frames, text, faces, and bodies (Aizawa et al., 2020), narra-

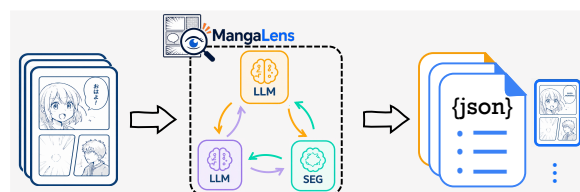


Figure 1: **Overview of MangaLens.** MangaLens combines LLM-based reasoning and segmentation-based visual grounding to support manga parsing subtasks and produce structured, inspectable JSON records.

tive reasoning over panels (Iyyer et al., 2017), and multi-task understanding benchmarks (Vivoli et al., 2024). However, downstream manga workflows often require structured, inspectable parsing results that connect text, visual regions, characters, and reading order, rather than isolated predictions from individual subtasks. For example, manga translation is not a text-only task: dialogue must be interpreted together with page layout, visual context, speaker identity, and narrative flow (Hinami et al., 2021). This motivates integrated tools that expose intermediate parsing outputs as reusable artifacts.

We present MangaLens, an integrated tool for structured manga parsing. As shown in Figure 1, MangaLens processes manga pages through collaborative interactions between LLMs and a segmentation model. The LLMs reason over page structure, narrative context, reading order, and relationships among manga elements, while the segmentation model grounds panels, characters, and speech bubbles. The system produces page-level JSON records and a book-level character record that are both machine-readable and human-verifiable.

MangaLens supports subtask-level visualization throughout the parsing workflow. After each parsing step is completed, the interface displays the corresponding outputs with aligned image evidence, including OCR results, panel narratives, character evidence, and other intermediate artifacts. Users

can inspect the final JSON records and correct parsing errors by referring to the aligned image regions. The resulting structured records can be reused as manga structural annotations and as inputs to downstream applications. This work makes the following contributions:

- **An integrated tool for structured manga parsing.** We present MangaLens, a tool that combines LLM-based reasoning with segmentation-based visual grounding to support manga parsing subtasks and produce page-level JSON records and a book-level character record.
- **Subtask-level visualization and correction.** MangaLens visualizes the outputs of individual parsing steps with aligned image evidence, allowing users to inspect OCR results, speaker links, character evidence, page summaries, and JSON records, and to correct parsing errors.
- **Evaluation of structured parsing outputs.** We evaluate MangaLens across major parsing subtasks and include manga translation as an additional downstream use case, showing that the tool produces reliable structured parsing outputs and reusable records.

2 Related Work

Manga and Comic Datasets. Annotated datasets and benchmarks have defined many core subtasks for manga and comic parsing and understanding. Manga109 provides annotations for frames, text regions, character faces, and bodies across 109 manga books (Aizawa et al., 2020). MangaSeg further augments Manga109 with pixel-level segmentation annotations for manga elements such as frames, speech bubble regions, text/dialog regions, onomatopoeia, character faces, and character bodies (Xie et al., 2025). The COMICS dataset studies narrative reasoning across panels and shows the importance of combining visual and textual context (Iyyer et al., 2017). CoMix further broadens comic understanding to multiple tasks, including object detection, speaker identification, reading order, character re-identification, character naming, and dialogue generation (Vivoli et al., 2024). These resources define important task settings, while MangaLens focuses on integrating their outputs into structured, inspectable records.

Manga transcription and narrative modeling. The Magi line of work also studies manga parsing and transcription. The first Magi system (Sachdeva and Zisserman, 2024) introduced a unified model

for manga transcription that detects panels, text boxes, and character boxes, clusters characters, associates dialogue with speakers, and recovers reading order. A later version (Sachdeva et al., 2024) extended this direction to chapter-wide manga transcription with character names, emphasizing narrative consistency and cross-page character identity. More recently, a further extension (Sachdeva and Zisserman, 2025) used comic parsing outputs with vision-language models to generate literary narratives from comics. Unlike prior work that primarily focuses on manga transcription or narrative generation, MangaLens is designed as an integrated tool for structured manga parsing, exposing subtask outputs as inspectable, correctable, and reusable page-level and book-level JSON records.

Vision-language reasoning and visual grounding. Recent vision-language models have shown strong capabilities in document understanding, structured extraction, visual localization, and multi-image reasoning (Bai et al., 2025b,a). In parallel, promptable segmentation models have progressed from image-level segmentation with SAM (Kirillov et al., 2023) to image-and-video segmentation with SAM2 (Ravi et al., 2024) and concept-prompted detection, segmentation, and tracking with SAM3 (Carion et al., 2025). MangaLens builds on these developments by combining LLM-based structural reasoning with segmentation-based grounding to support structured manga parsing.

3 System Overview

MangaLens is built around a collaborative parsing loop between an LLM group and a segmentation model. The LLM group consists of two LLMs. We use A to denote this group, which proposes structural, textual, and reasoning outputs, and S to denote the segmentation model, which grounds visual targets in the image. As an integrated tool for structured manga parsing, MangaLens organizes parsing into output-producing operations for panel layout, visual entities, page-local character entities, text/dialogue information, and cross-page character identities. These operations produce page-level JSON records and a book-level character record, and further support identity-aware page story summaries. All structured parsing outputs are produced using only these three models, without additional task-specific models or character re-identification models. Figure 2 illustrates the overall workflow of MangaLens.

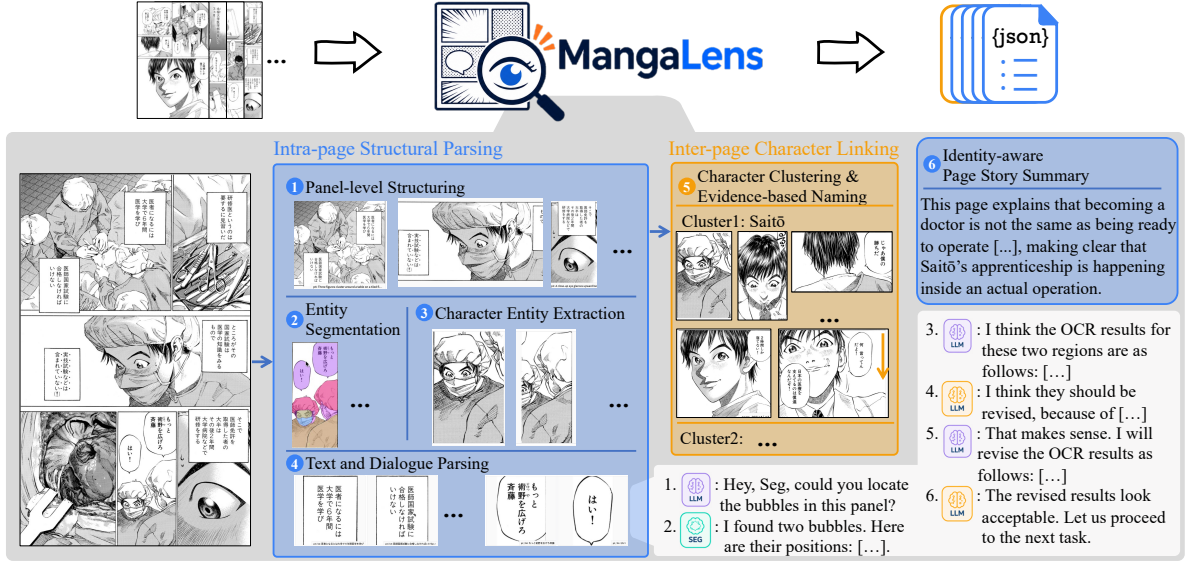


Figure 2: **Overview of the MangaLens workflow.** MangaLens first performs **Intra-page Structural Parsing** to convert each page into a structured JSON record with panels, grounded visual entities, OCR text, reading order, speaker links, and panel-level narratives. It then performs **Inter-page Character Linking** to aggregate page-local character entities into a book-level character record with recurring identities and evidence-based names or fallback labels, which are used to generate identity-consistent page story summaries. The inset illustrates how *A* and *S* collaborate to identify speech bubble regions and perform OCR within the detected regions.¹

Given a sequence of manga page images $\mathcal{X} = \{x_i\}_{i=1}^N$, MangaLens first constructs page-level structured records and then aggregates them for book-level character reasoning:

$$R_i = F_{\text{intra}}(x_i), \quad \mathcal{R} = \{R_i\}_{i=1}^N, \\ \mathcal{C} = F_{\text{inter}}(\mathcal{X}, \mathcal{R}).$$

Here, R_i denotes the structured record of page x_i , including panels, visual entities, OCR text, reading order, speaker links, and panel-level narratives. Each R_i is serialized as a page-level JSON record, $\mathcal{R}$ denotes all page-level records, and $\mathcal{C}$ is serialized as a book-level JSON record for recurring character identities and supporting evidence. In the first stage, *Intra-page Structural Parsing*, MangaLens constructs R_i for each page independently. In the second stage, *Inter-page Character Linking*, MangaLens aggregates $\mathcal{R}$ with the original page images $\mathcal{X}$ to produce the book-level character record $\mathcal{C}$. The resulting character record is then used to generate identity-aware page story summaries, which are written back to the corresponding R_i . The detailed JSON structures are provided in Appendix A.

¹All manga images shown in this paper are from the first chapter of *Give My Regards to Black Jack*, authored by SHUHO SATO, and are taken from the official freely available version: <https://densho810.com/free/>.

3.1 Intra-page Structural Parsing

The first stage parses each page independently into R_i through four steps: panel-level structuring, entity segmentation, character entity extraction, and text & dialogue parsing. These steps progressively construct the panel layout, grounded visual entities, page-local character entities, and text/dialogue information of each page.

For each parsing task τ , the LLM group first produces a candidate output from the current visual input and task context:

$$\hat{R}_\tau = A_\tau(z_\tau, c_\tau), \\ R_\tau = \left\{ \{r \in \hat{R}_\tau \mid S(z_\tau, r) \neq \emptyset\}, \tau \in T_g, \right. \\ \left. \hat{R}_\tau, \tau \in T_d. \right.$$

Here, z_τ is the visual input for task τ , such as the full page image, a panel crop, or a speech bubble crop; c_τ is the task context; and $\hat{R}_\tau$ is proposed by A_τ . For grounding tasks $\tau \in T_g$, MangaLens keeps only candidates localized by S with a non-empty mask or box. For non-grounding tasks $\tau \in T_d$, such as OCR, reading-order inference, and speaker attribution, the LLM-generated result is directly stored as R_τ .

Step 1: Panel-level Structuring. This step covers panel detection, panel reading-order inference, and panel narrative generation. Following prior work that treats panel detection and reading order

as core subtasks (Sachdeva and Zisserman, 2024), S detects panel regions, while A infers their reading order and generates a short narrative for each panel. This step produces panel regions, ordered panels, and panel-level narratives in R_i .

Step 2: Entity Segmentation. This step covers semantic target enumeration and entity grounding, with speech bubble detection handled as part of grounding. Within each detected panel, A enumerates visible targets such as characters, speech bubble regions, body parts, clothing, and other visual entities, following prior manga segmentation work (Xie et al., 2025). These targets are converted into segmentation prompts and grounded by S into instance-level boxes and masks, which are stored as visual entities in R_i .

Step 3: Character Entity Extraction. This step extracts and organizes page-local character entities from the grounded visual entities. Each entity stores its panel, box, mask, crop, and page-local identifier. These entities provide candidate speakers for speech bubble–speaker linking in Step 4 and candidate character appearances for cross-page character linking in Step 5.

Step 4: Text and Dialogue Parsing. This step covers speech bubble OCR, OCR for selected non-bubble text regions, dialogue-order inference, and speech bubble–speaker linking. Following manga speaker detection settings that require text ordering and speaker attribution (Li et al., 2024), OCR is applied to speech bubble regions and selected non-bubble text regions such as captions and signs, excluding stylized sound effects. The LLM group A infers dialogue order and speaker links using speech bubble positions, page-local character entities, OCR text, panel narratives, and neighboring panel context. The resulting OCR text, dialogue order, and speaker links, and speech bubble spatial metadata are inserted into R_i .

3.2 Inter-page Character Linking

The second stage implements the inter-page mapping:

$$\mathcal{C} = F_{\text{inter}}(\mathcal{X}, \mathcal{R}).$$

Unlike intra-page parsing, this stage treats character reasoning as a book-level problem because character appearances may be partial and name cues may be sparse or absent.

Step 5: Character Clustering and Evidence-based Naming. This step covers cross-page character clustering and evidence-based naming. Following chapter-level manga transcription work on cross-page character identity and name consistency (Sachdeva et al., 2024), MangaLens collects page-local character entities from $\mathcal{R}$ and pairs them with visual crops from $\mathcal{X}$. The LLM group A clusters visually and contextually compatible candidates into recurring character identities using visual evidence, OCR text, dialogue content, and speaker links. Each cluster is assigned either an inferred name or a stable fallback label, and $\mathcal{C}$ stores the recurring identities, names or fallback labels, and supporting evidence.

3.3 Identity-aware Page Story Summary

After obtaining $\mathcal{C}$, MangaLens generates an identity-consistent story summary for each page:

$$y_i = A_{\text{sum}}(x_i, R_i, \mathcal{C}).$$

Here, A_{sum} denotes the summarization instance of A . Inspired by manga-to-prose work that uses manga parsing outputs to generate literary narratives (Sachdeva and Zisserman, 2025), the summary y_i is conditioned on the original page image x_i , the page-level record R_i , and the book-level character record $\mathcal{C}$, and is stored as a structured field in R_i .

4 Implementation and User Interface

MangaLens is implemented in Python. Its front-end interface is built with Gradio (Abid et al., 2019) and gradio-sam-prompter², allowing users to run the parsing pipeline, inspect intermediate visual results, and reuse saved parsing outputs. The structured parsing pipeline uses three models: GPT-5.5³ and Qwen3.6-27B (Qwen Team, 2026) for LLM-based reasoning, and SAM3 for visual grounding. Qwen3.6-27B is served with vLLM (Kwon et al., 2023) to accelerate generation. SAM3 is used to localize manga elements such as panels, speech bubble regions and other visual entities.

The interface is designed to simplify operation while preserving structured results for inspection, correction, and reuse. As shown in Figure 3(a), users only need to provide a folder of single-page manga images, with filenames ordered according

²<https://github.com/hysts/gradio-sam-prompter>

³<https://developers.openai.com/api/docs/models/gpt-5.5>

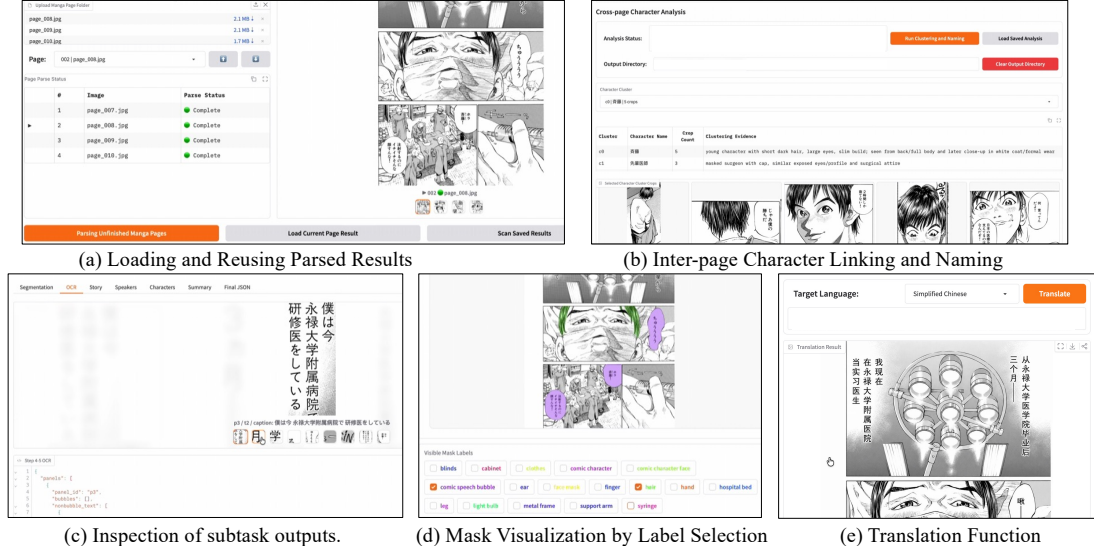


Figure 3: **User Interface of MangaLens.** The interface shows MangaLens-generated annotations, masks, structured outputs, translated manga pages, and character-clustering crop results.

to the page sequence. They can then start *Intra-page Structural Parsing* with a single button click and monitor the parsing state through a status table. The parsing outputs are saved as structured JSON records and can be reloaded later: users can scan for existing parsed JSON files and inspect saved records without rerunning the full pipeline.

Although the main outputs of MangaLens are JSON records, the interface presents them through synchronized textual and visual views. Users can inspect structured JSON fields together with their corresponding image evidence, such as entity masks, speech bubble regions, text-region crops, and OCR results. The interface organizes subtask outputs into separate tabs, allowing users to examine OCR results, story descriptions, speaker links, character evidence, page summaries, and the final structured JSON, as shown in Figure 3(c). Users can also visualize different masks by selecting entity labels, as shown in Figure 3(d). This helps users verify parsing outputs and correct OCR results, speaker links, entity labels, or other JSON fields by referring to aligned image evidence. In this sense, MangaLens can also serve as a lightweight tool for inspecting and correcting manga structural annotations.

After all pages in the selected folder have been parsed, users can run *Inter-page Character Linking* by clicking the *Run Clustering and Naming* button, as shown in Figure 3(b). This function takes the saved page-level records $\mathcal{R}$ as input, clusters page-local character entities across pages, and assigns evidence-based names or fallback labels. The re-

sulting book-level character record $\mathcal{C}$ is saved as a structured JSON record and can be reused together with the page-level records.

Finally, MangaLens includes a manga translation function as an additional downstream application of the page-level JSON records, as shown in Figure 3(e). This function is not part of the structured parsing pipeline, but is used to test whether the parsing records provide useful constraints for downstream manga translation. Given a target language, the LLM component translates the parsed OCR text using the reading order, speaker-related information, and page context stored in R_i , while GPT-image-2⁴ uses the translated text and speech bubble regions to generate an edited translated page.

5 Evaluation

We design the evaluation around two data sources: the first chapter of the Japanese version of *Give My Regards to Black Jack* for human evaluation and Manga109 for partial task-level testing. The *Give My Regards to Black Jack* chapter contains 60 high-quality sequential pages with diverse layouts, speech bubbles, recurring characters, dialogue-heavy scenes, and cross-page context, making it suitable for assessing the full structured parsing pipeline. Manga109 is used to test selected tasks on a standard manga dataset and compare with prior results where available.

We evaluate seven major parsing outputs: panel detection, panel reading order, speech bubble de-

⁴<https://developers.openai.com/api/docs/models/gpt-image-2>

Task	Black Jack		Manga109				Metric
	MangaLens	Metric	MangaLens	Magiv3	MangaLMM	MangaSeg	
Panel detection	0.996	BAP@0.5	0.959	0.92	—	0.954	BAP@0.5
Panel reading order	0.950	Acc.	0.832	—	—	—	Acc.
Speech bubble detection	0.941	BAP@0.5	0.917	—	—	0.914	BAP@0.5
OCR	0.058	CER↓	0.747	—	0.715	—	Hmean(NED)
Speech bubble–speaker linking	0.839	Acc.	0.889	—	—	—	Acc.
Dialogue order	0.841	Acc.	0.738	—	—	—	Acc.
Inter-page character linking	0.828	Acc.	0.587	0.678	—	—	AMI

Table 1: **Evaluation results for structured parsing outputs.** Black Jack results are based on human evaluation of the first chapter of *Give My Regards to Black Jack*. Manga109 results are partial task-level tests on selected subtasks, with comparisons to prior work where available. Bold numbers indicate the best result among available Manga109 results for the same task. Higher values are better for all metrics except CER, where lower values are better. “—” indicates that the work does not support or report results for that task. Magiv3 (Sachdeva and Zisserman, 2025), MangaLMM (Baek et al., 2026), and MangaSeg (Xie et al., 2025) are included as reference results.

tection, OCR, speech bubble–speaker linking, dialogue order, and inter-page character linking. On *Give My Regards to Black Jack*, two annotators familiar with Japanese manga manually evaluate these outputs, resolving disagreements through discussion. On Manga109, we evaluate the same task types where applicable, but only some have corresponding prior results, so the comparisons in Table 1 are necessarily partial. For sequence-based outputs, such as panel reading order and dialogue order, a page is counted as correct only when the whole predicted sequence is correct. The OCR metric differs across the two data sources: *Give My Regards to Black Jack* uses CER for speech bubble text, while Manga109 uses a NED-weighted end-to-end Hmean over annotated text regions. Other outputs, such as general entity segmentation, character extraction, panel narratives, page story summaries, and character naming, are not evaluated separately because they require specialized annotation criteria and substantially more manual labeling.

The Manga109 results are summarized in Table 1. For panel and speech bubble detection, MangaLens achieves strong results through SAM3-based prompt grounding and uses no task-specific detection training. For OCR, MangaLens may return larger text-containing regions than the corresponding Manga109 annotations, making direct IoU-based matching sensitive to the difference in region granularity. We therefore use coverage-based region matching and then compute the NED-weighted end-to-end Hmean, as detailed in Appendix B. Under this evaluation setting, MangaLens achieves an Hmean of 0.747 on Manga109 without manga-specific training. Prior work reports that the best-performing general-purpose LMM evaluated without manga-specific fine-tuning achieves an end-to-end Hmean of only

0.9% (Baek et al., 2026), illustrating the difficulty of end-to-end manga OCR for general-purpose LMMs.

On *Give My Regards to Black Jack*, MangaLens performs reliably on visually grounded tasks, achieving high scores for panel and speech bubble detection and a low CER for speech bubble OCR. Tasks that require richer visual context remain more challenging, including speech bubble–speaker linking, dialogue ordering, and inter-page character linking. In particular, inter-page character linking errors often occur when overlapping characters in the same frame are merged by crop-based evidence, suggesting that segmentation-based character regions may improve clustering.

Overall, MangaLens produces reliable structured outputs for layout and text-centered parsing, while speaker attribution, dialogue ordering, and cross-page character identity remain challenging. We further apply the parsed records to downstream manga translation; qualitative human evaluation shows that the MangaLens-assisted pipeline improves readability and faithfulness and reduces hallucinated content compared with direct Image-to-Image translation, as reported in Appendix C.

6 Conclusion

We presented MangaLens, an integrated tool for structured manga parsing. MangaLens combines LLM-based reasoning with segmentation-based visual grounding to produce inspectable and correctable page-level and book-level JSON records. By exposing subtask outputs with aligned visual evidence, the system supports verification, correction, and reuse of parsing results across layout, speech bubble parsing, OCR, speaker linking, and cross-page character reasoning. Evaluation results show that MangaLens produces reliable structured

parsing outputs across major parsing tasks. Additional human evaluation on manga translation further demonstrates the usefulness of the parsed records for downstream applications.

7 Limitations

MangaLens is a prototype system, and its performance depends on the quality of both LLM-based reasoning and segmentation-based grounding. Irregular panel layouts, dense typography, stylized fonts, occluded characters, chibi drawings, and silhouette-like appearances can still lead to parsing errors. Because no existing dataset fully matches the structured records and subtask coverage produced by MangaLens, our evaluation mainly relies on human assessment of major parsing outputs, supplemented by limited tests on existing benchmarks. These benchmark tests are necessarily partial, as they cover only selected subtasks rather than the full structured parsing pipeline.

Ethics and Broader Impact

MangaLens processes manga images, which may be copyrighted; users should only process works for which they have appropriate rights or permissions. The manga images used in this paper are taken from the official freely available version of *Give My Regards to Black Jack*, authored by SHUHO SATO, under the work’s reuse terms.

References

- Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.
- 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.
- Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. 2026. MangaVQA and MangaLMM: A benchmark and specialized model for multimodal manga understanding. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5349–5370, Rabat, Morocco. Association for Computational Linguistics.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-v1 technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-v1 technical report. *arXiv preprint arXiv:2502.13923*.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, and 1 others. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.
- Neil Cohn. 2014. The architecture of visual narrative comprehension: The interaction of narrative structure and page layout in understanding comics. *Frontiers in Psychology*, 5:680.
- Neil Cohn. 2016. A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, 146:304–323.
- Grand View Research. 2025. Manga market size, share & trends analysis report. <https://www.grandviewresearch.com/industry-analysis/manga-market-report>. Accessed: 2026-07-01.
- Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12998–13008.
- Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry 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.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Yingxuan Li, Kiyoharu Aizawa, and Yusuke Matsui. 2024. Manga109dialog: A large-scale dialogue dataset for comics speaker detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Ragav Sachdeva, Gyungin Shin, and Andrew Zisserman. 2024. Tails tell tales: Chapter-wide manga transcriptions with character names. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2053–2069.
- Ragav Sachdeva and Andrew Zisserman. 2024. The manga whisperer: Automatically generating transcriptions for comics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12967–12976.
- Ragav Sachdeva and Andrew Zisserman. 2025. From panels to prose: Generating literary narratives from comics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21864–21873.
- Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. 2024. Comix: a comprehensive benchmark for multi-task comic understanding. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA. Curran Associates Inc.
- Minshan Xie, Jian Lin, Hanyuan Liu, Chengze Li, and Tien-Tsin Wong. 2025. Advancing manga analysis: Comprehensive segmentation annotations for the manga109 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8869–8878.

A JSON Output Structures

MangaLens produces two formal JSON outputs: a page-level record for each page and a book-level character record for the whole input sequence. The page-level record corresponds to R_i in Section 3, while the book-level character record corresponds to $\mathcal{C}$. We use `string`, `number`, `boolean`, and `null` to denote primitive JSON types, and `T[]` to denote an array of elements of type `T`. Bounding boxes are represented as `[x1, y1, x2, y2]`.

A.1 Page-level Record

Each page-level record is serialized as a `page_record.json` file, with its main fields shown in Figure 4 (top). It stores the structured parsing result of a single page, including panel layout, grounded visual entities, OCR results, dialogue structure, local character entities, and the final identity-aware page story summary. Identifiers such as `panel_id`, `entity_id`, and `bubble_id` are local to the page.

The page-level record links visual, textual, and narrative information through local identifiers. For example, `panel_id` connects panels with entities, masks, bubbles, and dialogue entries; `mask_id` connects grounded visual objects to segmentation masks; and `speaker_entity_ids` connects dialogue bubbles to page-local character entities. The field `page_summary` stores the final identity-aware story summary written back after book-level character linking.

A.2 Book-level Character Record

The book-level character record is serialized as `character_clusters.json`, with its main fields shown in Figure 4 (bottom). It aggregates page-local character entities from all page-level records into globally addressable candidates, whose identifiers combine the page identifier with the local entity identifier, such as `page_007:e2`.

In this record, `clusters` groups compatible global entities into recurring character identities, `unlinked_entity_ids` stores candidates that are not safely assigned to any cluster, and dialogue and background OCR provide textual evidence for evidence-based naming.

B Manga109 OCR Evaluation Details

Manga109-style OCR evaluation first matches predicted text regions to annotated text boxes and

then computes recognition scores. This protocol assumes that predictions are text-region-level boxes. However, MangaLens is designed for structured manga parsing and often produces larger text-containing regions, such as speech bubbles. As shown in Figure 5, Manga109 annotations tightly enclose glyph regions, while MangaLens predictions may cover the same text with a larger surrounding region. Directly applying IoU matching would therefore penalize correct text coverage because of box granularity mismatch.

To align the matching step with the MangaLens output granularity, we use ground-truth text coverage for MangaLens OCR evaluation on Manga109:

$$\text{Cov}(G, P) = \frac{|G \cap P|}{|G|},$$

where G is a Manga109 text box and P is a MangaLens predicted region. If the predicted region covers the annotated text, this score remains high even when P is larger than G .

C Qualitative Evaluation on Downstream Manga Translation

We further evaluate MangaLens through a downstream manga translation task. This study examines whether the structured records produced by MangaLens provide useful constraints for translation. We compare a direct Image-to-Image baseline, where GPT-image-2 translates the original page directly, with a MangaLens-assisted pipeline, where translation uses OCR results, speech bubble regions, reading order, speaker-related information, and page context from the parsed records.

We evaluate Chinese and English translations of the first chapter through human assessment. Due to potential copyright concerns, we report qualitative evaluation results without reproducing the translated page images. As shown in Table 2, direct Image-to-Image translation performs poorly for Chinese and shows low faithfulness for both target languages, including hallucinated content such as the severe case on p.47. Although its English outputs can be fluent, they often fail to preserve the original content. In contrast, the MangaLens-assisted pipeline achieves good readability and higher faithfulness in both Chinese and English. Its remaining errors mainly involve difficult Japanese expressions or proper-noun renderings, such as keeping or mistranslating “永大出”. These results suggest that structured manga parsing can improve

```

{
  "image_path": string,
  "source_image": {
    "width": number,
    "height": number,
    "mode": string,
    "sha256_rgb": string
  },
  "panel_order": string[],
  "page_bubble_reading_order": string[],
  "panels": [
    {
      "panel_id": string,
      "order_index": number,
      "box": [number, number, number, number],
      "story": string,
      "objects": [
        {
          "obj_id": string,
          "label": string,
          "box": [number, number, number, number],
          "mask_id": string,
          "entity_id": string | null,
          "bubble_id": string | null
        }
      ],
      "entities": [
        {
          "entity_id": string,
          "type": string,
          "name_hint": string,
          "panel_id": string,
          "box": [number, number, number, number]
        }
      ],
      "bubbles": [
        {
          "bubble_id": string,
          "box": [number, number, number, number],
          "mask_id": string,
          "panel_id": string,
          "label": string,
          "box": [number, number, number, number],
          "mask_encoding": string
        }
      ],
      "dialogue": [
        {
          "bubble_id": string,
          "text": string,
          "speaker_entity_ids": string[],
          "panel_id": string,
          "order_in_panel": number,
          "order_in_page": number
        }
      ],
      "page_summary": string,
      "page_summary_source": string
    }
  ],
  "status": string,
  "scope": string,
  "page_json_paths": string[],
  "entities": [
    {
      "global_entity_id": string,
      "page_id": string,
      "panel_id": string,
      "local_entity_id": string,
      "box": [number, number, number, number],
      "crop_path": string,
      "page_json": string
    }
  ],
  "clusters": [
    {
      "cluster_id": string,
      "entity_ids": string[],
      "name_hint": string,
      "confidence": number,
      "naming_confidence": number,
      "naming_evidence": string[],
      "note": string
    }
  ],
  "unlinked_entity_ids": string[],
  "dialogue": [
    {
      "page_id": string,
      "panel_id": string,
      "bubble_id": string,
      "text": string,
      "speaker_global_entity_ids": string[],
      "order_in_page": number
    }
  ],
  "background_ocr": [
    {
      "page_id": string,
      "panel_id": string,
      "text_id": string,
      "role": string,
      "text": string
    }
  ],
  "page_summaries": {
    "<page_id>": string
  }
}

```

Figure 4: **JSON output schemas.** Top: main fields of the page-level record serialized as `page_record.json`. Bottom: main fields of the book-level character record serialized as `character_clusters.json`.



Figure 5: **OCR box granularity mismatch.** Green boxes indicate MangaLens predictions, and blue boxes indicate Manga109 annotations. Courtesy of Yoshi Masako.

downstream translation by constraining the process with OCR text, speech bubble regions, reading order, and speaker-related context.

Target	Method	Read.	Faith.	Main observations
Chinese	Direct I2I	Poor	Low	Numerous translation errors; uncomfortable to read. Severe hallucination on p. 47.
Chinese	MangaLens	Good	Higher	Generally smooth; a few JP-CN mixed words remain. E.g., “永大出” kept on p. 17, though later correctly rendered as “永大学毕业的”.
English	Direct I2I	Good	Low	Fluent English but many unfaithful translations; some pages introduce hallucinated story lines.
English	MangaLens	Good	Higher	Mostly natural; main issues are Japanese proper-noun renderings, e.g., “永大出” translated as “Nagaide” instead of “Eidai grad”.

Table 2: **Qualitative human evaluation of downstream manga translation.** We compare direct Image-to-Image translation with the MangaLens-assisted pipeline for Chinese and English versions of the first chapter of *Give My Regards to Black Jack*. “Read.” denotes readability, and “Faith.” denotes faithfulness.