Developing a LLM-Driven Multi-Agent Framework for Multimodal Translation

Nicholas Jumaoas
njumaoas@ucsd.edu

Fiona Jiang
fejiang@ucsd.edu

Kristina Wu
mew013@ucsd.edu

Hao Zhang
haozhang@ucsd.edu

Abstract

This research presents a multi-agent framework for multimodal translation, focusing on manga and literary content translation. The framework integrates specialized agents that handle different aspects of the translation process, from text detection to typesetting. By combining OCR-based text extraction, LLM-driven translation with visual context awareness, and automated typesetting, the system aims to produce high-quality translations while maintaining style consistency and cultural nuances. Initial results demonstrate the framework's potential for processing raw content into reader-ready translations, suggesting promising applications for expanding this approach to other media types. The research contributes to advancing automated translation systems by addressing the challenges of context preservation and visual-textual integration in multimodal content.

Website: TBA
Code: https://github.com/nljumaoas/dsc180b

1 Introduction

The increasing global demand for translated content, particularly in multimedia formats like manga, comic book, and literary works, has highlighted significant challenges in maintaining both efficiency and quality in translation processes. Traditional manual translation workflows, which rely heavily on human translators, editors, and typesetters, often struggle with scalability and consistency when handling large volumes of content that combines both textual and visual elements. This challenge becomes particularly acute in the context of manga translation, where cultural nuances, visual context, and textual accuracy must be carefully balanced by skilled professionals, leading to substantial time investments and increased production costs.

Recent advances in artificial intelligence, particularly in Large Language Models (LLMs) and multi-agent systems, present new opportunities to address these challenges. While current machine translation systems have made significant strides in handling pure text, the complexity of multimodal content—where meaning is conveyed through both text and images—remains a substantial challenge. The integration of visual context into translation decisions, the maintenance of consistency across large documents, and the preservation of stylistic elements are areas where traditional automated approaches often fall short.

This research proposes a novel approach to multimodal translation by developing a framework that mirrors the organizational structure and workflow of professional translation companies. By leveraging the strengths of both artificial intelligence and traditional translation methodologies, our system aims to create a more robust and adaptable translation pipeline. The framework incorporates specialized agents handling distinct aspects of the translation process, from initial text detection to final typesetting, working in coordination to produce high-quality translations that maintain both accuracy and cultural sensitivity.

1.1 Literature Review

Machine translation (MT) has evolved significantly in recent years, driven by advancements in neural networks and the increasing availability of large parallel corpora. Traditional MT systems, such as those based on recurrent neural networks (RNNs) and attention mechanisms, have been significantly outperformed by transformer-based architectures (Vaswani et al. 2023). These transformer models have become the backbone of state-of-the-art translation systems, enabling more accurate and fluent translations across diverse languages and domains. However, despite these advancements, challenges remain, particularly when translating context-heavy content such as images with embedded text, where visual context plays a crucial role in ensuring accurate translations.

Hinami et al. (2021) utilize visual context in an automatic machine translation framework, as well as providing annotated, professionally translated pages as a base to be used in future research. Lippmann et al. (2024) adapt a similar framework to replace conventional MT with a LLM augmented by visual context. They also investigate the effect of varying context lengths on translation quality, concluding that simply incorporating a LLM does not result

in improvement beyond the page level.

Recent research has explored multi-agent systems as a means to further enhance translation quality. One notable work in this domain is TransAgent (Wu et al. 2024), which introduces a multi-agent reinforcement learning framework for literary translation. In TransAgent, multiple translation agents engage in a cooperative process, leveraging each other's strengths to produce higher-quality translations. This approach has shown promise in handling the nuanced and stylistic aspects of literary texts, which require a deep understanding of language subtleties and cultural context.

Building on the foundations laid by TransAgent, our research aims to adapt the multi-agent framework to the domain of image-based translation. Unlike pure literary translation, image-based translation involves interpreting textual content within visual contexts, necessitating a model that can effectively integrate multimodal information. The combination of textual and visual data poses unique challenges, as the meaning of the text can be heavily influenced by the accompanying imagery. To address these challenges, our approach incorporates visual context into the multi-agent translation process, allowing agents to debate and refine their translations based on both linguistic and visual cues.

The efficacy of multi-agent systems in improving task performance has been further supported (Liang et al. 2024), who demonstrated that frameworks involving agent debates lead to significant performance enhancements in various natural language processing (NLP) tasks. This finding underscores the potential of multi-agent interactions in refining translation outputs, as agents can collaboratively resolve ambiguities and inconsistencies.

Our research seeks to bridge these two strands of literature—multi-agent systems and multimodal translation—by developing a novel framework that adapts the cooperative agent-based approach to context-heavy image translation. By enabling agents to debate and integrate visual context into their translation decisions, we aim to achieve superior translation performance, particularly in scenarios where traditional single-agent or purely text-based models fall short.

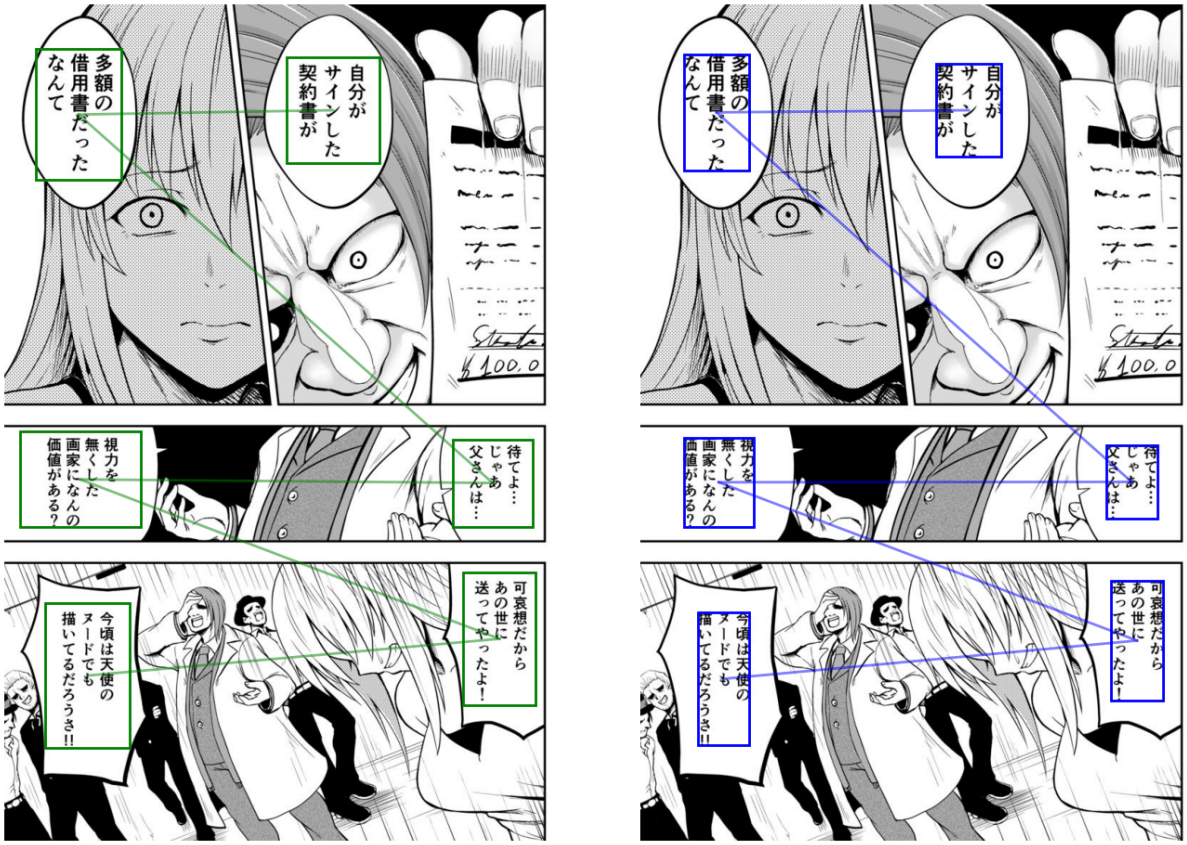
2 Methods

2.1 Page Processing

In order to match benchmarks such as OpenMantra (Hinami et al. 2021) for evaluation purposes as well as provide a consistent and accessible format for later stages in the pipeline, it is most logical for the input to include an accompanying dictionary with annotations detailing information about manga page elements and extracted text. Although the idea of using a LLM-driven page-processing agent is entertained as an additional part of the multi-agent framework, the procedure is quite linear and as such does not require the flexibility provided by LLM integration; as such, a processing pipeline is implemented instead, taking inspiration from prior work on automatic manga translation (Lippmann et al. 2024).

The initial input is manga pages in the form of images; while it has only been tested on dig-

ital media, this approach should also work for physical pages. The input is first passed into a text segmentation model that utilizes unconstrained text detection, which accounts for many manga-specific edge cases, including variation in dialogue element shapes, non-text elements included within speech bubbles, and text outside the dialogue elements, although the latter is notably recognized as difficult and inconsistent since it has to account for non-white backgrounds for text, which is standard within the speech bubbles (Gobbo and Herrera 2020). The model produces a text mask that is then clustered using the `pyclustering` library's implementation of the OPTICS algorithm (Novikov 2019), resulting in boxes that can be used to identify text elements, primarily speech bubbles.



(a) OpenMantra dataset

(b) Page-processing pipeline

Figure 1: A visual comparison of the page-processing pipeline to the OpenMantra baseline; although the identified fields and order are the same, the processed boxes fit tighter to the text, facilitating text replacement in the typesetting stage.

To determine reading order, specific functionalities of the Magi model (Sachdeva and Zisserman 2024) are used to detect and order panel elements before being combined with the text elements identified in the previous steps, producing a sorted list of text boxes by analyzing the relative positions of the text clusters within the ordered panels. Using the `Pillow` library (Clark 2015), each of these boxes are used to produce a cropped image that isolates the identified text from the surrounding image, allowing it to be easily processed for text extraction. This is performed by the MangaOCR model, which is designed to be ro-

bust against scenarios specific to manga, such as text overlaid over images or accompanied by furigana, wide varieties in font style and size, and different reading orientations (Budyś 2024).

The end result of this stage is formatted into a dictionary that adds to each processed image ordered lists of identified panel and text elements, specifically their coordinates and side lengths. The latter set of elements also include the corresponding text extracted by MangaOCR. This format is chosen to align with the format of OpenMantra's page annotations, both for the purpose of evaluating the page-processing output and for providing a baseline to be used by later stages. Figure 1 showcases a visual comparison of the page-processing output to the annotations provided by OpenMantra.

2.2 Multi-agent Context-aware Translation Design

In the second stage of our methodology, we introduce a multi-agent context-aware translation system designed to enhance the accuracy and coherence of translations, particularly for context-heavy image-based content such as manga, comic books, and illustrated narratives. This stage leverages a combination of Visual Language Models (VLMs) and a multi-agent collaborative framework to ensure that translations are both linguistically accurate and contextually relevant.

2.2.1 Visual Context Analysis

The process begins with the application of a Visual Language Model (VLM) to analyze the visual context of the page. The VLM extracts semantic information from the images, identifying key elements such as characters, settings, and visual cues that contribute to the narrative context. This ensures that the subsequent translation process is informed by the visual nuances integral to the meaning of the text.

2.2.2 Multi-agent Translation Framework

Following the visual context analysis, the translation process is handled by a multi-agent system comprising three specialized agents. Each agent is designed with a distinct profile and specialization to address different aspects of the translation task. The Linguistic Specialist Agent focuses on grammatical and syntactical correctness, ensuring that the output adheres to the linguistic norms of the target language. The Cultural Context Specialist Agent emphasizes cultural nuances and idiomatic expressions, adapting the translation to fit the cultural context of the target audience while preserving the original intent. The Visual Context Specialist Agent integrates the visual information extracted by the VLM, ensuring that the translation aligns with the visual elements and narrative flow depicted in the images.

2.2.3 Incorporation of Historical Context

For image series with a continuous storyline, such as manga or comic books, the system incorporates historical context into the translation process. This involves tracking narrative elements across multiple pages to maintain consistency in character names, plot development, and thematic elements. By referencing previous translations and narrative developments, the system ensures coherence throughout the entire series.

2.2.4 Agent Collaboration and Verification

The three agents engage in a round-robin style group chat to debate and verify translation outputs. Each agent generates an initial translation based on their specialization, followed by a review of each other's translations, providing critiques and suggestions for improvement. This iterative process allows for the identification and resolution of discrepancies. Through multiple rounds of debate, the agents converge on a final translation that balances linguistic accuracy, cultural appropriateness, and visual coherence. The consensus translation undergoes a final verification phase where the agents collectively assess the translation against the visual and historical context to ensure holistic consistency.

This multi-agent, context-aware translation design leverages the strengths of specialized agents and visual analysis to produce high-quality translations that are both accurate and contextually appropriate, addressing the complexities inherent in translating visually rich narratives.

2.3 Typesetting Design

In the third stage of our methodology, we introduce a typesetting system designed and built from scratch to seamlessly integrate translated text into manga panels while preserving the original visual layout. This stage focuses on ensuring that the translations are accurately formatted, dynamically sized, and positioned to maintain aesthetic consistency and readability within the artwork.

2.3.1 Precise Text Wrapping and Dynamic Font Adjustment

The translated text is wrapped dynamically based on the pixel width of each line, measured using the PIL library. This custom-built approach ensures precise line breaks, particularly for languages with varying word lengths. To handle text overflow, the system adjusts the font size recursively until the text fits within the speech bubble without clipping or overflow. This mechanism guarantees readability while preserving visual clarity.

2.3.2 Layout-Aware Centering and Text Replacement

The pipeline, developed from the ground up, centers the wrapped text both vertically and horizontally within the speech bubble for a balanced and appealing layout. Before adding the translated text, the original content is cleared by drawing a white rectangle over the text area. The translated text is then inserted with a stroke effect, enhancing visibility against various background colors commonly found in manga panels.

2.3.3 Output Flexibility and Customization

After processing, the modified image is saved as the output, with options for customization, such as different fonts, styles, and stroke effects. The system's built-from-scratch design allows for high flexibility, making it adaptable to various typesetting needs across different genres and ensuring smooth integration into translation workflows.

3 Results

4 Discussion

5 Conclusion

References

Budyś, Maciej. 2024. "Manga OCR." [Link]

Clark, Alex. 2015. "Pillow (PIL Fork) Documentation." [Link]

Gobbo, Julián Del, and Rosana Matuk Herrera. 2020. "Unconstrained Text Detection in Manga." [Link]

Hinami, Ryota, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. "Towards Fully Automated Manga Translation." [Link]

Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate." [Link]

Lippmann, Philip, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. 2024. "Context-Informed Machine Translation of Manga using Multimodal Large Language Models." [Link]

Novikov, Andrei V. 2019. "PyClustering: Data Mining Library." *Journal of Open Source Software* 4(36), p. 1230. [Link]

Sachdeva, Ragav, and Andrew Zisserman. 2024. "The Manga Whisperer: Automatically Generating Transcriptions for Comics." [Link]

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. "Attention Is All You Need." [Link]

Wu, Minghao, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. "(Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts." [Link]

Appendices

A.1 Project Proposal

A copy of the project proposal that our group turned in at the end of DSC180A can be found here; however, please note that while we have remained within the same field of multi-agent systems, our current sub-niche of multimodal translation is not discussed in that proposal. To uphold the spirit of providing background for how this project came to be, links to our topic proposal and progress update slideshows are provided below:

- Topic Proposal
- Week 3 Progress
- Week 4 Progress
- Week 5 Progress

A.2 Training Details

A.3 Additional Figures

A.4 Additional Tables

B Contributions

- Nicholas Jumaoas:
  - Conducted initial research on primary sources (TransAgents, Multimodal Manga Translation) as well as secondary sources to verify prior work, benchmarks, and other logistics.
  - Adapted page processing pipeline to match benchmark standards and fit the needs of other pipeline stages, as well as accommodate evaluation methods such as latency and text/element recognition accuracy.
  - Contributed to literature review and page processing sections of report checkpoint.
- Feiyang Jiang:
  - Conducted holistic literature review on relevant previous research and products in multi-agent LLM translation of text with images task to ensure novelty of the proposed problem.
  - Established working framework of translation stage of the pipeline incorporating visual context from VLM and multi-agent LLM debate framework for enhancing translation quality. Organized stages into pipeline.
  - Contributed to multi-agent setup in literature review and translation stage method explanation in the report writeup.
- Kristina Wu:
  - Conducted literature review on diffusion and impainting models relevant to the last stage of the project pipeline.
  - Established reproducible working system of the typesetting stage, incorporating text removal, embedding, and fusion in a sample manga/comic page.
  - Contributed to writing the abstract, general introduction to literature review and typesetting stage method explanation in the report writeup.