Multimodal Context Informed Machine Translation of Manga Using LLMs

Master's Thesis

Konrad Skublicki

Multimodal Context Informed Machine Translation of Manga Using LLMs

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Konrad Skublicki
born in Krosno, Poland

TU Delft
Web Information Systems Research Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Multimodal Context Informed Machine Translation of Manga Using LLMs

Author: Konrad Skublicki
Student id: 5863813

Abstract

Large language models have achieved breakthroughs in many natural language processing tasks. One of their main appeals is the ability to tackle problems that lack sufficient training data to create a dedicated solution. Manga translation is one such task, a still budding and underdeveloped field, that at the same time deals with a problem even more complex than standard translation. One of its biggest challenges is successfully incorporating the visual modality in translation, to resolve ambiguities.

Recently, a new type of LLMs – multimodal LLMs – has emerged and showed potential in understanding visual symbolism in narrative pieces like memes and comics. In this work, we investigate whether these models could be useful in the field of manga translation, ultimately helping manga authors and publishers make their works available to wider audiences. Specifically, we evaluate a number of methods based on GPT-4-Turbo for text only translation, image informed translation and volume-level translation. We perform both automatic and human evaluation. Moreover, we contribute new evaluation data – the first parallel Japanese-Polish manga translation benchmark.

Our findings show that our proposed methods are able to achieve state of the art results for English, and set a new standard for Polish. We conclude that while this is not a sufficient replacement for a professional human translator, it could help speed up the translation process or be used as a learning aid. The code is available on GitHub [1].

Thesis Committee:

Chair:	Associate Prof. Dr. Christoph Lofi, Faculty EEMCS, TU Delft
Daily supervisor:	Assistant Prof. Dr. Jie Yang, Faculty EEMCS, TU Delft
Committee Member:	Associate Prof. Dr. Cynthia Liem, Faculty EEMCS, TU Delft

[1] https://github.com/sqbly/multimodal-context-manga-translation

Preface

This thesis represents the culmination of my Master's studies at TU Delft. It is a result of extensive research, diligent work and support and guidance of many individuals, whom I would like to thank sincerely.

First and foremost, I am deeply grateful to my supervising professor, Dr. Jie Yang. You encouraged me to find a subject I was deeply interested in, which made the entire process extremely satisfying and rewarding. Your feedback guided me on this academic journey and made sure I stayed on the right path. Thank you for believing in my potential and working with me.

I would also like to sincerely thank Philip Lippmann for being with me every step of the way. Your contributions, whether through bold ideas, insightful discussions, or critical reviews, have significantly enhanced the quality of this thesis. Your patience in addressing my countless questions have made this journey both enriching and enjoyable. The collaborative spirit and intellectual exchange that you brought to this project were invaluable, and I have learned immensely from working alongside you.

Next, I would like extend my thanks to the employees at Mantra Inc. for collaborating with us. I would like to especially thank Joshua Tanner for being so actively involved and helping arrange the human evaluation, which I would like to thank Elizabeth Lillie Ramos for carrying out. I would also like to express gratitude to Hiroto Kaino and his co-authors for kindly sharing their translation outputs with us.

I would like to take this opportunity to thank my Japanese teacher, Anna Koike-Kamińska, for her expert input on the manga translation industry and linguistic advice.

Last but not least, I am also grateful to my fellow graduate students, friends and family, whose moral support has been a source of encouragement throughout this journey. I would like to specifically mention my friend Felix Kaubek, who was my closest partner throughout my time at TU Delft. Your willingness to engage in thoughtful discussions and provide constructive feedback has been deeply appreciated. Couldn't have done it without you buddy.

Konrad Skublicki
Delft, the Netherlands
June 24, 2024

Contents

List of Figures

Chapter 1

Introduction

Manga – a Japanese style of comics – is a multi-billion dollar industry, making up a large portion of the domestic publishing market in Japan, as well as 73% of the international copyright-related trade of books and comics [101, 18, 127]. However, manga translation is a difficult and time-consuming process, leading to a large discrepancy between the number of manga titles being published in Japan and those available overseas [2, 47, 40, 121]. What is more, even if a title is available overseas, it is usually only accessible to speakers of major languages such as English, French and Spanish. This has given raise to the phenomenon of *scanlation* - a grassroots movement of fans scanning, translating and unofficially distributing manga in their own languages, to popularize otherwise unavailable titles, often without any financial incentive [71, 22, 26]. This has been an ongoing issue for publishers, as the availability of those, in essence pirated, copies of their copyrighted titles makes overseas readers less likely to purchase an official translation should one be created, decreasing the incentive for investing in translation in the first place, completing the vicious circle. Moreover, in the case of the Polish manga industry, researchers have reported that readers will often criticize the official Polish translators for their stylistic choices, if they differed from the scanlated versions they read before, and therefore consider closer to original [26].

As the root cause of all those issues is the speed at which official translations could feasibly be made, machine translation (MT) has been proposed as a potential solution [40, 47, 128]. MT is field that has been established 70 years ago, but is still rapidly developing. With the popularization of neural machine translation (NMT), the past decade has seen many breakthroughs [39, 91]. However, even though translation from Japanese has always been challenging [48, 106], the problem of manga translation poses even more unique challenges, that SOTA general purpose NMT solutions – usually trained for sentence-to-sentence translation on datasets made up of news and scientific articles – are not equipped to deal with. These include resolution of ambiguities using visual semantic information extracted from hand-drawn images, literary translation, maintaining consistency with previous translations and handling sentences split across multiple speech bubbles.

Research into manga-specific NMT methods has started, but is currently very limited, focusing on only one translation direction: Japanese-English, and held back by the lack of substantial parallel corpora for all language pairs [40, 47]. Only one manga translation dataset has been published so far, containing 214 Japanese pages (1593 sentences) with translations into English and Chinese, effectively making it an evaluation dataset. All previously proposed models were trained on a private Japanese-English dataset, that is not shareable due to copyright issues and has no known counterpart in other languages. Moreover, only one of the previously proposed methods has dealt with informing translation with visual context, by means of detecting a small number of predefined labels, but achieving inconclusive results [40].

Aware of these limitations, this thesis aims to approach the manga translation problem with large language model (LLM) empowered machine translation. LLMs have brought about breakthroughs in many NLP tasks, including MT [65, 39]. LLM translation has even shown promising results in literary translation, a task that is traditionally difficult for most general-purpose translation systems [48]. Moreover, with the release of the GPT-4 model [83], multimodal machine translation with LLMs has become a possibility [66]. Using models like that, we theorize it might be possible to circumvent the problem of obtaining large parallel manga corpora for each language separately, and create a system that is able to handle many translation directions at once.

1.1 Research Question

The specific research question this thesis will aim to answer is:

How effective are multimodal LLMs in translation scenarios where incorporating multimodal context is necessary to bridge the information gap between typologically distant languages?

To help in answering that, the following sub-questions will be investigated:

1. How big is the impact of visual context on the quality of translation produced by state of the art multimodal large language models?
2. How well can such models scale up to translating long narratives?
3. How consistent is the performance across different language pairs?

To achieve that, we will evaluate the out-of-the-box performance of the state of the art multimodal multilingual large language model GPT-4-Turbo [83] on the existing Japanese-English manga datasets as well as on a new Japanese-Polish dataset that we will create for this purpose. This specific translation direction has been chosen in line with previous machine translation research [48, 106], as it presents some unique challenges. For example, Japanese is known for zero pronouns often taking place of subjects and direct objects in sentences. This, coupled with the fact that this language also lacks morphological markers of person, gender and number, means that outside information – or broader context – is required to correctly resolve the zero pronoun. When translating to a language like English, which also does not have such morphological markers, this lack of information will not impact the translation. However, in Polish, the antecedent of a zero pronoun is usually communicated through grammatical information marked on other elements of the sentence (e.g. conjugating a verb by number, person or gender). Consequently, obtaining a correct translation is often impossible without resolving these ambiguities [106]. As such, translating manga into Polish should serve as a more demanding benchmark for measuring the understanding of visual context, as the missing information about, for example, the gender of the speaker, should be apparent from the content of the page.

1.2 Contributions

The contributions of this thesis paper are as follows:

- A LLM-based multimodal manga translation system that achieves state of the art results on translation to English, and can serve as a baseline for all other languages
- The first publicly available automatic manga translation evaluation suite [1]
- An annotation set of 400 professionally translated manga pages (3705 sentences), that together with scans included in the Manga-109 dataset [30] make up the first ever parallel Japanese-Polish manga translation benchmark dataset [2]

1.3 Outline

The rest of the paper is divided as follows: Chapter 2 gives the overview of the related work in the fields of manga translation, multimodal machine translation, LLM machine translation and manga processing. Chapter 3 explains the specific translation approaches that will be evaluated in this paper, and proposes a manga processing pipeline for page-to-translation workflow, although it is not the main focus of this work. Chapter 4 presents the data that the proposed solutions will be tested on, describes the the baseline methods and explains the metrics that will be used to compare them. Chapter 5 presents the results and chapter 6 discusses the outcomes, as well as limitations and ethical considerations. Chapter 7 gives closing remarks and lists promising directions for future work. Lastly, appendix A contains all the prompts used for evaluation.

[1] Available at `https://github.com/sqbly/multimodal-context-manga-translation`

[2] Available on request from the author

Chapter 2

Related Work

While machine translation is a well established field, manga translation is a small and under-researched niche. This chapter discusses previous attempts at machine translating manga. Following that, relevant research into multimodal machine translation and machine translation using large language models is reviewed. Lastly, other manga processing technologies are introduced.

2.1 Machine Translation of Manga

2.1.1 Towards Fully Automated Manga Translation (2021)

Arguably the most important work regarding manga translation is *Towards Fully Automated Manga Translation (2021)* by Hinami et al. [40]. It is the first study focused on creating a neural machine translation system specifically for manga. The authors argue that an accurate translation is impossible to obtain without incorporating context beyond just the text to be translated. They focus on three types of additional context information: scene context, reading order and visual information.

The additional context information for all three types of context is produced before the translation by separate methods proposed by the authors. For scene detection, the authors trained the Faster R-CNN detector [93] with the Manga109 dataset [30, 74, 4]. For estimating the reading order, the authors manually design an algorithm based on general rules of manga reading. For visual information, the authors use the illustration2vec [99] model to detect 512 pre-defined semantic tags for each scene. Text detection is also solved by a vision model designed and trained by the authors. None of of the models or code for these methods is shared publicly.

Incorporating such obtained context, the authors propose two new, and in total evaluate three NMT models based on the Transformer (big) model [112]. The first of the evaluated methods is the *2+2 translation* [110], which involves giving the model the previous sentence as extra context for the sentence that is being translated. The second method, proposed by the authors themselves, is *Scene-based translation*, which is a generalization of the *2+2 translation*. In this case, the scenes correspond to lines appearing in speech bubbles within the same panel. The model is trained to translate all utterances from one scene at once. The last model, *Scene-based translation with visual feature*, is an extension of the second method, where the lines from each scene are prepended with tokens corresponding to predicted semantic tags. All of the models are trained on a parallel Manga Corpus of over 840 thousand manga pages in both English and Japanese, that has been automatically collected by the authors, but has not been made public for legal reasons. The authors do however publish a smaller evaluation dataset, that will be used in this work and discussed in more detail in chapter 4.

For evaluation, the authors choose three baselines that use sentences as translation units: Google translate, and two NMT models created by the authors – trained on the OpenSubtitles18 [60] and previously discussed Manga Corpus respectively. Results show that on the automatic metrics, BLEU [87] in this case, the best method is in fact, the *Sentence-NMT* baseline, that does not incorporate any extra contextual information. However, human evaluation shows that the new methods incorporating context, namely *Scene-based translation* and *Scene-based translation with visual feature* are statistically significantly better than that baseline. The authors argue that this is because a fluent translation will often swap the order of sentences around, which is penalized by BLEU, but actually favoured by humans. They continue on to suggesting that BLEU might not be suitable for evaluating manga translations. The results also showed that incorporating visual context actually decreased the scores, although not significantly. Results of manual investigation suggest, that this was most likely due to incorrectly detected visual features. The authors point to better image encoding methods and models considering context longer than one scene as promising areas for future work.

While both Hinami et al. [40] and the present work investigate the incorporation of multi-modal context for machine translation of manga, there are several key differences. Unlike [40], this work investigates systems, where the visual and text modalities are processed by the same model. Furthermore, we investigate translation units many times longer than one scene. Lastly, our proposed method uses a general purpose multilingual model, that does not need to be trained for each language pair separately, and does not require collecting large amounts of parallel manga data.

2.1.2 Utilizing Longer Context than Speech Bubbles in Automated Manga Translation (2024)

Kaino et al. [47] build upon the work of Hinami et al. by investigating the effects of providing the NMT model with broader contextual information than one scene. They investigate two types of context: the text of up to 5 previous scenes (*Scene-aware NMT)* and bibliographic information, such as the genre and author of the work that is being translated (*Attribute-aware NMT*). Unlike [40], this work does not consider the visual context whatsoever.

The NMT models used in this work are based on the Transformer (big) model [112] using fairseq [84]. The data used for training and evaluation is the Manga Corpus created in [40]. This was possible, because both papers share some of the authors. As baselines, Sentence-NMT, as well as *2+2 translation* and *Scene-based translation* [40] are used.

The authors report that both of their methods achieve improvement over the baselines, and the improvement is even greater, when the methods are combined. However, the improvement in BLEU score is much more pronounced than the improvement in the other, more sophisticated, metric they report – xCOMET [32]. Evaluation of the effect of different number of previous scenes being incorporated reveals that best performance is achieved when incorporating only the scene immediately before the one being translated. Out of the different types of manga metadata, only the publisher information did not improve performance, which the authors attribute to the fact that one publisher usually publishes mangas from many different authors and genres. Multimodal MT and multilingual MT models that have been pre-trained on a large corpus are suggested as promising areas of future work. As such, the current work could be viewed as a natural continuation of the work done by Kaino et al. The presented methods are making use of both multilingual and multimodal models, extending the context beyond one page and are evaluated on more than one language.

2.1.3 Towards a direct Japanese-Polish machine translation system (2017)

The work of Świeczkowska [106] is a notable mention in this section. The author discusses the challenges involved with Japanese-Polish machine translation and points out that (at the time) there were no satisfying solutions to the problem, nor were there any parallel Japanese-Polish language corpora. The author then goes on to argue that manga would be a great source of data to train a Japanese-Polish NMT system on, as there are many high quality translations of manga available on the polish market. A theoretical NMT system is then discussed, one that would incorporate both the text modality and a visual component that would provide the model with descriptions of the relevant pages and panels. However, as far as we have been able to establish, none of these plans were ever carried out. On that note, the present work is the first work we are aware of, that has made use of officially published manga to create a parallel Japanese-Polish dataset.

2.1.4 Other works

Most of the remaining works that deal with manga translation usually use Google Translate or DeepL as translation engines. The work of [103] tries to improve such obtained translations using a naïve post processing grammar correction method. They train a model to reoder translated sentences from Subject Object Verb to Subject Verb Object, to obtain a more natural translation, although the BLEU scores did not indicate improvement. The authors argue, that the translations, while possibly correct, were usually much different from the reference sentences, and as such could not get high BLEU scores regardless of their actual quality. The work of [121] is a comparative study of an online fan-based translating community Komikcast and Google Translate. They investigate the Japanese-Bahasa Melayu/Bahasa Indonesia language pair on the example of 'Golden Kamuy Vol. 1'. They note that, within the framework of Vinay and Darbelnet translation strategies [114], Google Translate tends to favor Literal Translation, while Komikcast uses Equivalence more frequently. The Equivalence strategy involves altering the words in the source to allow for a more natural sounding translation. Other works in this category usually focus on the computer vision challenges, such as detecting panel bounding boxes, speech bubbles, erasing the text, text recognition and putting the text back on the page, and never evaluate the translation whatsoever [61, 72].

2.2 Multimodal Machine Translation

In this section, we will discuss different fields of Multimodal Machine Translation in context of machine translation of manga. MT involving the audio modality, such as spoken language translation, also falls under this category, but is not discussed here, as it is not relevant to the task of manga translation.

2.2.1 Caption Translation

Multimodal Machine Translation is often synonymous with image caption translation or video caption translation. Within the framework proposed by [35], this task falls under the category of strongly-grounded data, as more likely than not, the visual modality is going to provide information that could improve the translation. One way of solving these problems, especially popular prior to the release of BERT and the GPT series, is training translation models from scratch. These in turn fall mostly under one of the two categories: systems that employ a separate visual encoder, the output of which was later fed into the translation model [27, 12, 13, 21, 6, 34, 56], and systems that train an end-to-end model [45, 38, 55]. However, recently there has been an increasing interest in using general-puspose models pre-trained on large datasets, that are later fine-tuned for the translation task [35, 100, 56]. A problem often seen by the researchers is that the models often learn to rely more on the text modality, and ignore the visual modality for the most part. To that end, incorporating probing signals in training [126] and contrastive adversarial training approaches [41] have been investigated.

However, while all of these methods seek to inform the translation using the visual modality, the way that an image description is related to the image it describes is much different, than the way that a manga dialogue relates to the manga page [40]. What is more, most datasets in this area of research focus on languages such as English, German and French, and oftentimes researchers only consider English as the source language in experiments.

2.2.2 Subtitle Translation

Movie subtitle translation in another popular task in Mulitmodal Machine Translation. Unlike video captions, which deal with descriptions of the events in the video, movie subtitle translation deals with translating the lines of dialogue spoken by the characters [57]. This task is considered weakly-grounded, as many many times characters would be speaking about something that is not shown on the screen, which in turn, again, results in the problem of models not making use of the visual modality as much [35]. Because of the discouraging results of baseline non-vision models performing just as well at this task, a lot of the research in this area is now focused not only on creating the best possible models [36], but primarily on creating datasets that would require the understanding of the visual modality in the first place [58, 57, 75] The work of Hinami et at. [40] uses an architecture similar to those explored for movie subtitle translation and test a baseline method trained on a subtitle dataset, so we consider it an accurate representative of this field.

2.2.3 Text Image Machine Translation

Although its focus is a little different, the Text Image Machine Translation (TIMT) task is somewhat similar to manga translation. TIMT is a cross-modal generation task somewhere in the middle between Text Image Recognition and standard Machine Translation, as the focus here is to, given an image with text on it, return an image with the text translated to another language, but still fitting into the image composition [67]. This is either by a cascade model or an end-to-end model [67, 68, 69, 70]. While the end goal of an end-to-end manga translation system would also be returning a page with translated text, given the original text, in TIMT the image context is always supposed to be discarded, which is not the case with manga translation.

2.3 Machine Translation using Large Language Models

Machine Translation with frozen Large Language Models is a rapidly developing field, that has emerged around 3 years ago. Researchers have evaluated bilingual models such as GLM-130B [122], as well as multilingual models, including, but not limited to: ChatGPT(gpt-3.5-turbo-1106) [120, 39, 115], GPT-3.5 (text-davinci-003) [77, 48, 39, 115], GPT-4 [66, 115], LLaMA2-7B [120], PaLM [113], BLOOM and BLOOMZ [77].

Most research so far has focused on sentence to sentence translation. Although the bilingual models were not as promising [122], PaLM was almost able to match SOTA in translation into English [113] and the GPT-3.5 models were able to surpass SOTA on some language pairs, such as German-English, Japanese-English and Chinese-English [39]. Moreover, although GPT-3.5 has been reported to perform comparatively poorly in translation to Arabic, GPT-4 has been able to solve that issue for the most part by employing a different tokenization technique [77].

Studies have shown that the choice of prompting templates has a big impact on the performance of the systems when applying the zero-shot approach, but choosing correct examples in a one-shot or few-shot fashion more consistently improves the performance and the stability of results. [122, 113, 77, 39]. To further capitalize on that, making use of Translation Memory (TM) to select better examples for the few-shot approach, using techniques such as fuzzy matching has been proposed recently and showed very promising results [77, 66].

Owing to the large context windows of modern LLMs, there has been a growing interest in evaluating their performance in document-level translation. This has taken a couple of different forms such as: translating a large document segment by segment using few-shot prompting to achieve consistency and evaluating the entire document at once [77, 122], translating entire paragraphs of up to nine sentences at once [48] and even longer texts [115, 39, 66]. GPT models, in particular the newer ones, have shown to perform well on this task, but evaluation of long text translation has been limited by the lack of adequate metrics. One of the recurring findings is that for document-level translation, the performance of one-shot and few-shot approaches does not differ significantly, likely because a longer text provides enough context on its own.

The interactive nature of LLMs has also given raise to research into tasks such as stylised translation [77, 66, 65] and culturally aware translation [120], both of which are very relevant to manga translation. While results so far have been good mostly on high-resource languages, these works have not evaluated the GPT-4 model, which has been shown to improve on such cases.

Lastly, the possibility of using vision-empowered Large Language Models for Multimodal translation has been pointed out recently by Lyu et al. (April 2024) [66], but the authors provided only one example for case study and no evaluation. Moreover, the provided example shows GPT-4 struggling to properly read text from the image, indicating that it is not yet a one-stop end-to-end solution. What needs to be done to make it so will be discussed in the present work.

2.4 Automatic Manga Processing

While the field of automatic manga translation is still budding, the research into other automatic methods for manga has a long history. Although manga is just a subset of comics, it has enough unique characteristics, such as being written in exclusively in Japanese, always coming in black and white, using a different reading order and page structure, as well as very specific sound effects, often appearing only in manga, that only research dedicated to manga specifically will be discussed in this section.

2.4.1 Text

The fundamental tasks in automatic manga processing are text detection and text recognition, which is often referred to as Optical Character Recognition (OCR). Although there exist general purpose text detection and OCR solutions, the text in manga exhibits uniquely high variation and necessity to take in contextual information, and as such general-purpose detection solutions are not suitable if they were not created with manga as one of the use-cases in mind [16, 119]. This also makes it unfeasible to simply apply OCR to an entire manga page, without detecting the specific text regions first. To address the problem of text detection, multiple methods have been proposed, such as multi-stage systems where at each stage a different CNN is used for a specific subtask, modified Faster R-CNN [16], training a classifier network to detect Japanese text at a pixel level [20, 50], training a network for text-block detection and using a hand-crafted algorithm to merge them [95], methods based on the YOLOv3 algorithm [118] and using generative adversarial networks [119]. Deciding which of those methods is the best is not obvious, as different methods make different assumptions as to which text is relevant. For example, [119] beats [16] on their evaluation dataset, which includes sound effects, something the previous method disregarded. Moreover, most of the methods mentioned do not share the code or models obtained, and as such it is difficult to benchmark them against one another. For OCR, popular choices include Tesseract OCR [1] and manga-ocr [2].

2.4.2 Page layout

A significant portion of manga related research focuses on detecting the ”building blocks” of a manga page, such as speech bubbles and story panels. Detecting speech bubbles is a task complementary to text detection, as usually finding a speech bubble enables the user to simply apply OCR to the entire bubble, to extract the lines being spoken. However, as text in manga is often found outside of speech bubbles as well, this is not a complete solution if one is interested in text outside of character utterances. Nevertheless, identifying the text bubbles and story panels on a page, is a necessary step to properly processing a page, as the reading order of a page depends on the relative location of these elements.

Older methods for speech bubble detection include rule based approaches, both text-independent [94] and text-aware [62], while newer works tend to rely on convolutional neural networks [23, 25, 98]. Similarly, methods for panel detection moved from heuristic-based algorithmic approaches [85, 42] to relying on CNNs trained from scratch [25, 98], fine-tuning pre-trained R-CNNs [96] and those combining both machine learning and heuristics [125]. The work of Ikuta et al. is a notable exception, proposing a pixel-wise comic panel segmentation with Mask R-CNN to solve the issue of elements crossing the panel boundaries (a stylistic device, often used by authors for emphasis, see Fig. 3.1) [43]. For reading order estimation, rule-based approaches have been proposed most often, consistently achieving accuracy over 90% [98, 40, 108, 124, 54]. All of those methods however, are wholly unequipped to deal with irregular cases, in which the reader needs to deduce the reading order based on the content of the page, not the layout of its elements. To address that, [59] propose a method that obtains the reading order using topological sort on a graph created based on sentence embeddings, achieving new SOTA.

2.4.3 Other tasks

Remaining object detection research for manga fall under one of the two categories: character/face detection [15] and multi-object detection, that uses one model to detect speech bubbles, panel frames and characters at once [80, 78, 98]. Making use of all that information, [98] propose a system for manga transcript generation, albeit only for translations of manga, featuring English text instead of the original Japanese, [104] propose a manga retrieval system that is able to look up specific scenes in manga based on descriptions provided by the user and [49] and [10] design tools for learning new languages while reading manga. Cross-language research that does not deal with translation includes the work of [14] who carry out sentiment analysis on different language versions of the same 4-koma mangas (a subset of comedy manga, in which each scene is always composed of four panels) to see if the results are comparable and the work of [33] who propose a method for generating missing sentences in manga for both English and French.

[1] https://github.com/tesseract-ocr

[2] https://github.com/kha-white/manga-ocr

Lastly, another notable manga processing task, especially important for manga translating is manga image inpainting. Natural image inpainting is a well known task of recreating missing parts of an image. In the case of manga, it is used to remove free-flowing text, that is placed on top of the drawing and not inside a speech bubble, which is very often the case for onomatopoeia for example. So far, there have been three works in this field, two that perform inpainting when given a mask [117, 81] and one work that presents an end-to-end method that first detects all the text on the page, and then removes it with inpainting [50]. As all of those works have been published at a similar time, no direct comparison between them has been carried out, but [117] and [81] were shown to outperform the baselines that were created for natural images, when evaluated on manga.

Chapter 3

Method

The process of end-to-end (page-to-page) manga translation, whether human or machine, can be divided into the following three steps:

1. Page processing – identifying elements of the manga page, text detection and reading order estimation;

2. Translating the text – knowing the transcript of the page, we relate the text back to the content of the page and find equivalents in the target language;

3. Typesetting – removing the source text from the page and inputting the translated text in a stylized font.

The main focus of this thesis is the second step - finding a correct translation for the source text. Nevertheless, for completeness we will present methods for both the first and the second steps of the pipeline, in sections 3.2 and 3.3 respectively. The third step will be omitted entirely, as it has been already discussed in previous research [40, 61, 72] and is not necessary for achieving a correct translation. However, this chapter will start with an introduction of terms used in regards to manga translation.

3.1 Terminology

The general structure of a manga page consists of multiple story panels, referred to simply as **panels**. A regular panel is an enclosed rectangular box, but that is not always the case – see Figure 3.1.

Panels often contain text, which can fall into one of the three types: enclosed, free flowing and sound effect. Enclosed text will usually be in some sort of container, usually a square box for for narrative text, or in a **speech bubble** for text spoken or thought by characters. Free flowing text will also convey characters thoughts or spoken words. It is often used when the line is spoken is in the background of the scene, and appears in clearly distinct clusters. Lastly, **sound effects**, or onomatopoeia, are usually words that are supposed to describe sounds that something makes, but it manga it also relates to states that characters are in (e.g. feeling proud or smiling) even if they are silent. The different types of text presentation can be seen on Figure 3.2. When talking about **lines** in this thesis, it will always mean the content of one speech bubble, narrative box or cluster of free-flowing text.

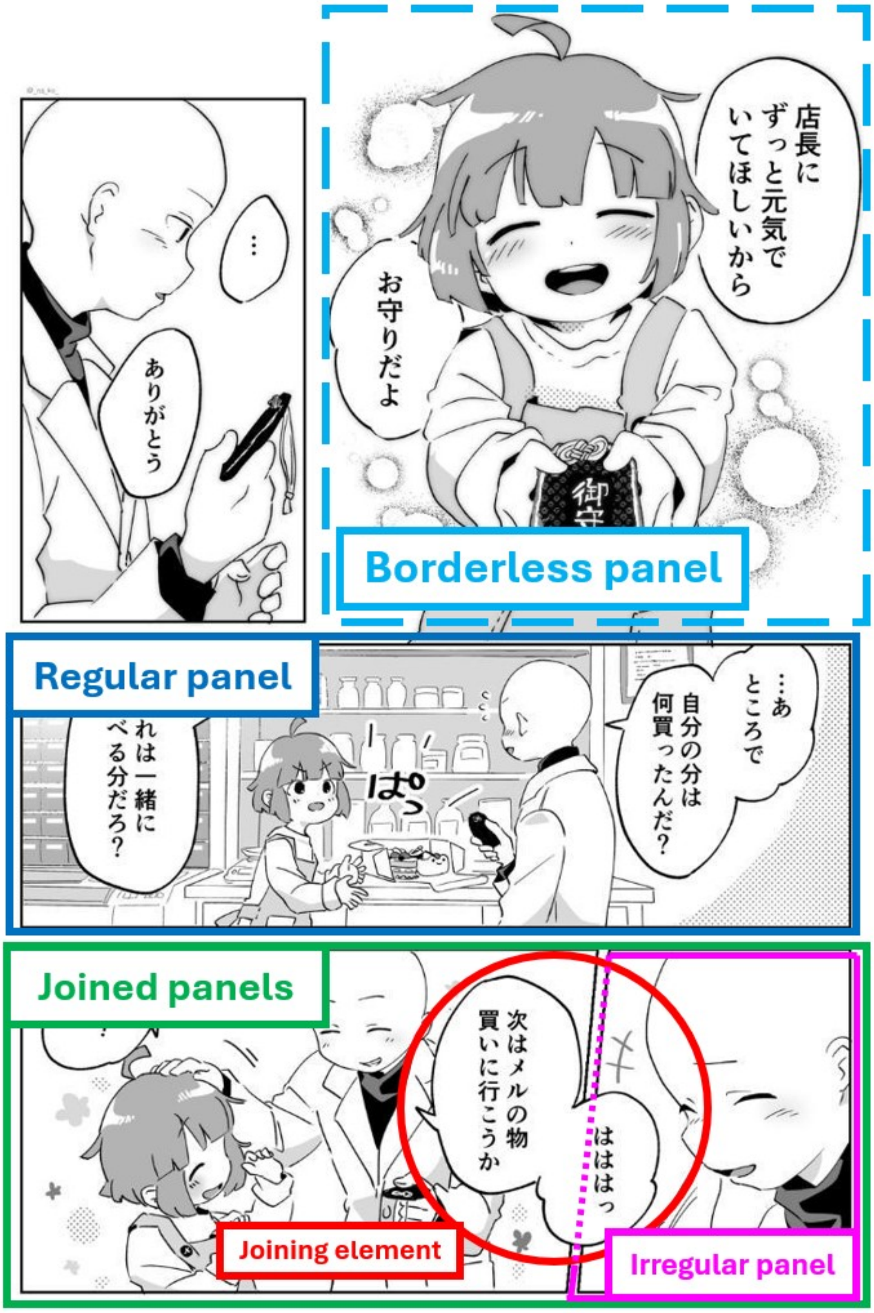


Figure 3.1: Examples of different types of panels seen in manga.

©Mitsuki Kuchitaka, from the OpenMantra dataset [40].

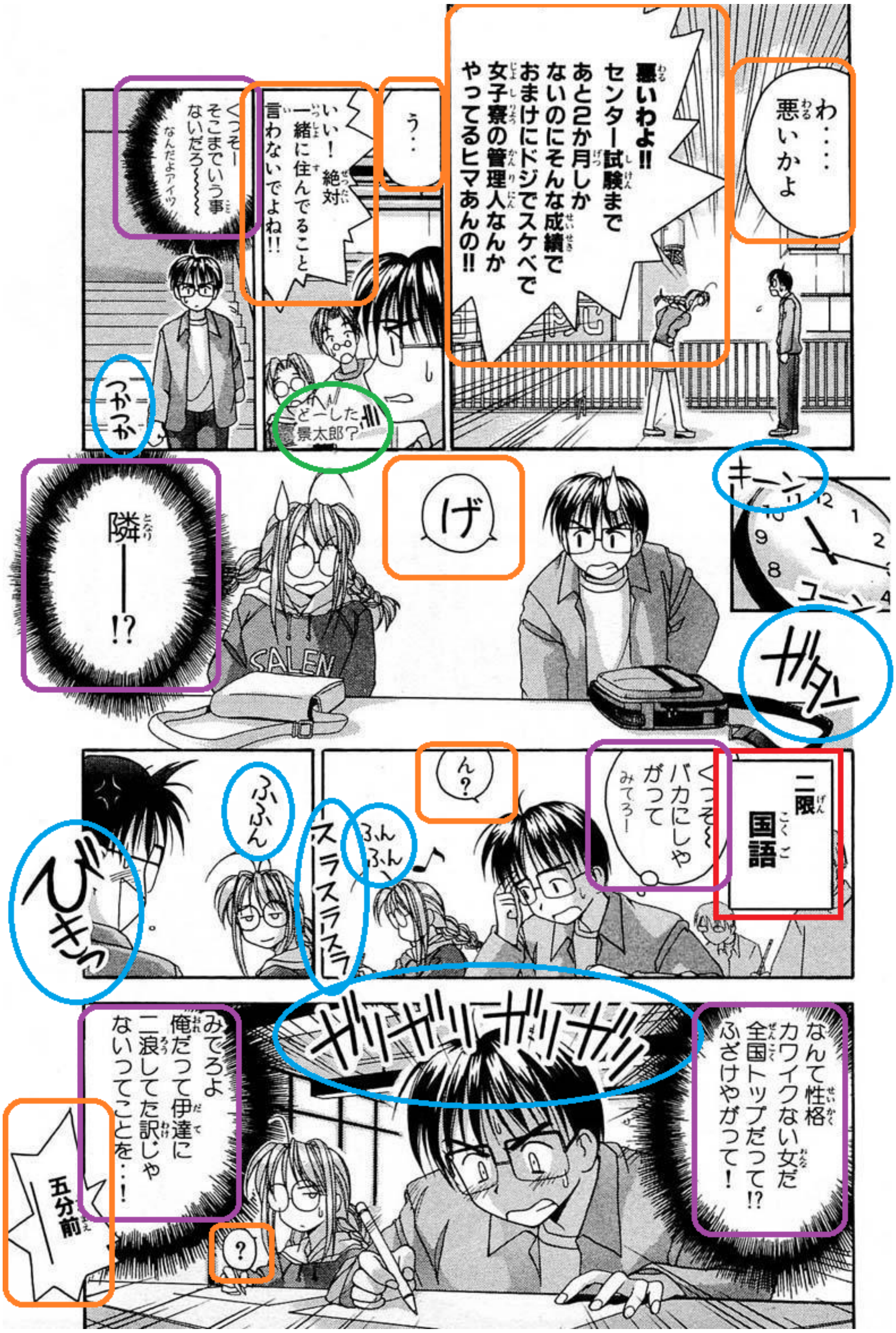


Figure 3.2: Examples of different types of text seen in manga.

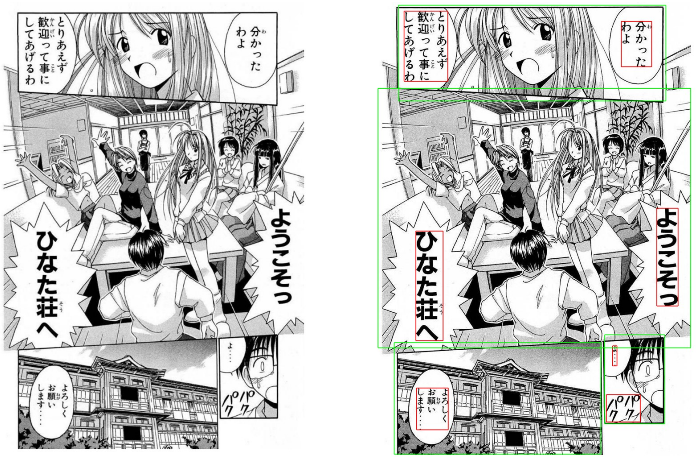
purple - speech bubbles (thoughts), **orange** - speech bubbles (conversation), **red** - narrative box, **green** - free flowing conversation text, **blue** - sound effects. Courtesy of Akamatsu Ken, ©Kodansha, from the Manga109-s dataset [30, 74, 4]

3.2 Page Processing

As mentioned at the beginning of this chapter, the first step of manga translation is identifying the elements on the page. In this section we will present an example manga page processing pipeline composed of methods proposed by previous research and publicly available manga tools.

For **text detection** we use the unconstrained method proposed by [20], to account for text that is not contained within speech bubbles – see Fig. 3.4 top right. However, to apply OCR to the detected text, we first need to group it into clusters belonging to the same utterance. In order to do that, we apply a method inspired by [95] – we use the OPTICS algorithm [5] to cluster the text, specifically the Python pyclustering library implementation [79] – see Fig. 3.4 bottom left. We then compute the bounding boxes of such obtained text clusters and discard those that are too small to contain text – see Fig. 3.4 bottom right. Lastly, we apply Manga OCR [1] for text recognition – see Fig. 3.3.

For **panel detection** and reading order estimation we use the Magi system by [98]. In theory, Magi is able to create a transcript of a manga page all on its own, but it was trained on translations of manga and is not well suited for Japanese text detection. As such, we only use some of its functionalities. A visualization of the page processing pipeline can be seen on Figure 3.3. First, we use the process described in the previous paragraph to detect text boxes. Then we use Magi to detect panels and estimate the reading order based on the relative locations of text boxes and panels. Lastly, we use Manga OCR for text extraction.



1. 分かったわよ
2. とりあえず歓迎って事にして あげるわ
3. ようこそっ
4. ひなた荘へ
5. よ. . .
6. パクパク
7. よろしくお願いします. . .

Figure 3.3: Page processing pipeline.

The reading order is estimated based on the relative location of the detected panels (**green**) and text boxes (**red**).
Courtesy of Akamatsu Ken, ©Kodansha, from the Manga109-s dataset [30, 74, 4]

[1] https://github.com/kha-white/manga-ocr

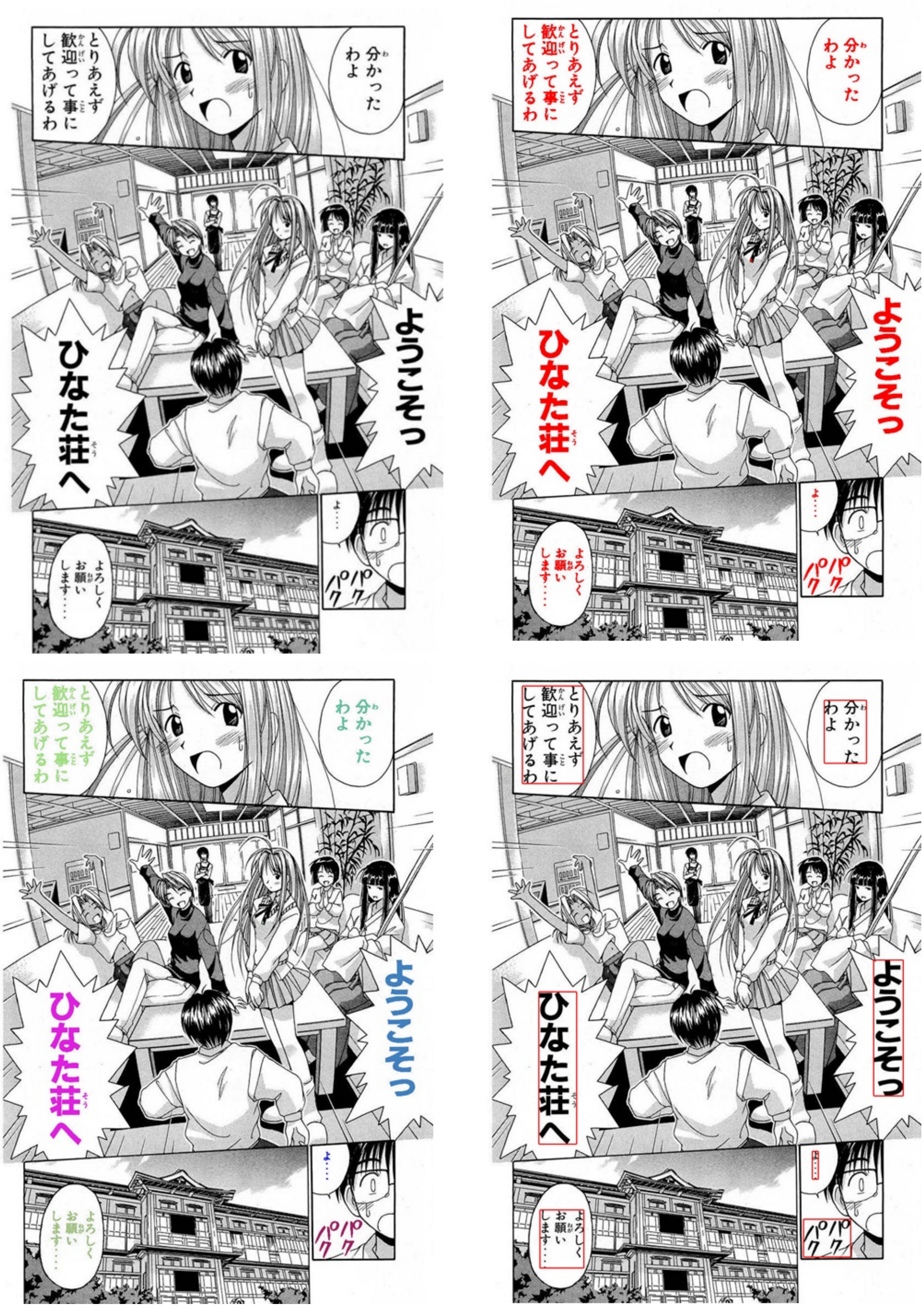


Figure 3.4: Stages of the text detection pipeline.

First, pixels belonging to letters are identified. Then, the pixels are clustered into utterances. Lastly, bounding boxes are computed.
Courtesy of Akamatsu Ken, ©Kodansha, from the Manga109-s dataset [30, 74, 4]

3.3 Translation

As previously stated, this thesis will follow in line of previous works [39, 77, 48, 66], and investigate the out-of-the-box performance of the most recent GPT model, GPT-4 Turbo [83]. The specific version used was `gpt-4-turbo-2024-04-09`, accessed through the OpenAI Python library [2].

This section will discuss the exact methods used for text only translation (3.3.1), image-informed page-to-page translation (3.3.2) and translation methods for larger translation units (3.3.3). All of the methods in this section were tested with both the *one-shot* and *few-shot* approaches. In case of the *few-shot* approach, 5 examples were always used, following [48]. In line with the findings of [122], the model was always prompted in English, regardless of the target language. The exact prompts used for all of those methods will be included in Appendix A. Table 3.1 shows a summary and comparison of all the proposed methods.

3.3.1 Text-only translation

To establish a point of reference, the first method investigated in this thesis is a simple line-by-line approach (`LBL`). This means that the model only received one line to translate and the examples, but no further context regarding the manga it is translating.

The second evaluated method is an analogous page-by-page approach (`PBP`), where the model is given all the lines from a given page, in the correct reading order. The findings of [48] showed that GPT-3.5 performed better when given the entire text, as opposed to having it translated line by line, as it was able to incorporate the broader context better. We expect to observe a similar effect on GPT-4-Turbo with `PBP` being strictly better than `LBL`.

3.3.2 Image-informed translation

In the ideal case, the model would be able to translate an entire manga page based on the image alone. This proved to be impossible with GPT-4-Turbo, which is in line with the disclaimer given in the Limitations section in the OpenAI vision documentation [3], that the model has limited capabilities when it comes to reading non-Latin scripts off of images. Instead, methods where the model is given the text to translate and images as additional context were investigated.

In experiments on the validation data, it was found that it is more effective to ask the model for an explanation on how the image informs the translation, to ensure the visual context is incorporated. As such, in all of the following methods, the models are tasked to return both the translation and the reasoning behind it, unless stated otherwise.

The first method is analogous to `LBL` – the model receives one line of text and the manga page it was taken from. We call it `LBL-VIS`, and will use the `VIS` component in the names of all methods that incorporate visual context. If the model is able to successfully incorporate visual context in translation, we expect `LBL-VIS` to be strictly better than `LBL`, as the latter

[2] `https://github.com/openai/openai-python`

[3] `https://platform.openai.com/docs/guides/vision/limitations`

receives much less information in comparison. It will also be interesting to compare with PBP, and see which is more informative – additional text or visuals.

The second method with visual context is PBP-VIS and as the name suggests it involves giving the model the entire text from one manga page and the page itself. In theory, this method combines the additional information from both PBP and LBL-VIS, making it the most informed method so far. Comparing the scores of all three of these methods should give us an insight into the size of the information overlap between the broader text context and the visual context.

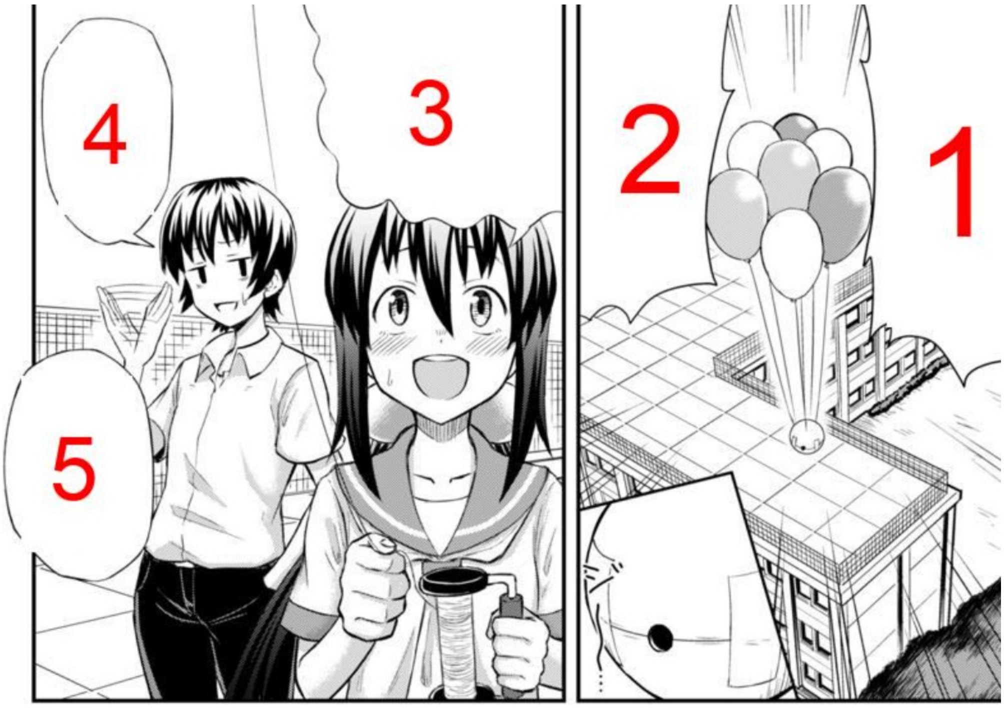


Figure 3.5: Fragment of a page annotated for the PBP-VIS-NUM method

©Mitsuki Kuchitaka, from the OpenMantra dataset [40]

The last method in this subsection aims to more directly address the issues that GPT models have with reading non-Latin scripts. The setup is the same as PBP-VIS, but the image of the manga page is modified. The contents of the speech bubbles are removed and replaced with numbers, indicating the reading order and corresponding to the list of Japanese lines the model is given to translate – see Fig. 3.5. In theory, this should enable the model to more easily locate the speech bubble and incorporate the context of the exact panel it was in. It would also provide additional help in determining the speaker of each line, although as can seen on the right side of Fig. 3.5 – the location itself is not enough reveal the speaker. The shape and the content of the bubble need to be analysed as well. We call this PBP-VIS-NUM, after the numbers placed inside the speech bubbles.

3.3.3 Long context translation

The end goal of manga translation is translating entire stories, sometimes multiple volumes long, in an internally consistent way. The rest of the methods we present will be designed to solve that task. However, due to the lack of suitable data, we limit the scope of this section to methods translating entire volumes at once, and leave multi-volume narratives for future work.

The first method we propose does not try to make use of the large context window size of GPT-4 and instead uses chain of density prompting (CoD) [3, 64] to maintain a rolling summary of the developments in the story so far. Apart from the image, the model is also given a summary of the story so far in the target language as additional context. It is then asked to, in addition to the translations, return the description of the events taking place on the page being translated, both in Japanese and the target language. A separate CoD module then prompts GPT to update the previous summary with the new developments, to achieve an even denser and updated summary. The summary is created in a 3-step process, in which each summary is the more concise version of the previous one, and the final outcome is saved. We call this method `VBP-VIS-COD`, as it translates the volume one page at a time (`VBP`), using visual context (`VIS`) and chain of density prompting (`COD`). We expect that the performance of this method will heavily depend on the quality of the summaries that are composed, as this is the first method in which the additional context is not obtained from an outside source.

The rest of the methods seek to achieve a more informed translation by providing the model with more pages at a time, making full use of the large context window of GPT-4. The first of these methods, for each page, provides the model with the previous and the next page as additional context and asks to translate all of them, in an effort to obtain a cohesive narrative across the entire story. We call this method `VBP-VIS-3P`. In theory, this method provides the model with more visual and more text information, so if `PBP-VIS` is better than `LBL`, then `VBP-VIS-3P` should be better than `PBP-VIS`.

So far, all the methods in this section continued the ”ask for an explanation about how the image informs the translation” approach. However, because of the token limit on the GPT API response, it is not possible to scale that approach to translating an entire 40+ page manga volume at once. As such, we adapt the `PARA-SENT` method from [48] to manga. Namely, we provide the model with the pages and lines from an entire manga volume, as well as all the translations so far, and ask it to translate only the next untranslated page, with explanations. This process is then repeated for every page in the volume. We call this method `VBP-VIS-ALL`, as the model is provided with all the available information and translates the volume page-by-page. Once again, this method provides the model with strictly more information than the previous method did.

The last of the evaluated methods, `VBV-VIS`, discards the explanation technique in favour of translating an entire manga volume in a single request. Similarly to `VBP-VIS-ALL` we provide the model with pages from an entire manga volume and all of the lines, and then instruct it to respond only with the translations, in an attempt to not exceed the completion token limit. If successful, this method would be many times more efficient than `VBP-VIS-ALL`, which made a request with a similar number of tokens, but as many times as there were pages in the volume. However, this method would also not scale as well as the previous ones, as for longer volumes it would hit the completion token limit sooner.

Method	Translation unit	Text context	Visual context
LBL	line	line	×
PBP	page	page	×
LBL-VIS	line	line	page
PBP-VIS	page	page	page
PBP-VIS-NUM	page	page	annotated page
VBP-VIS-COD	page	page + summary	page
VBP-VIS-3P	page	3 pages	3 pages
VBP-VIS-ALL	page	volume + tr. so far	volume
VBV-VIS	volume	volume	volume

Table 3.1: Overview of the proposed methods

Chapter 4

Experiments

Machine translation of manga is a small field in which there is very little publicly available evaluation data, and no set standard benchmarks, that researchers could use to compare their methods [40]. This chapter will introduce the decisions made in our experiments in light of this, and give reasoning behind each of those choices. Section 4.1 describes the data chosen (4.1.1) and created (4.1.2) for evaluation. Section 4.2 introduces the baseline methods (4.2.1), automatic metrics used to score the translation quality (4.2.2) and human evaluation methodology (4.2.3).

4.1 Data

There is a plethora of manga data that could be used for training and evaluation of machine translation systems [106, 98]. However, the inherent issue with manga data and manga translations is that due to its commercial nature, almost all of the data that exists is protected by copyright laws, making the matter even more complicated, when the original work is subject to Japanese copyright laws, but the translation is subject to both local and Japanese laws [101].

Previous manga related works dealt with this problem in different ways. Some resort to using ”private datasets” ([95, 40, 47]), others use the very few publicly available copyright-free manga, accepting the trade-off of having to work with unlabelled data ([103]). In [98], the authors decide to only work with translated manga – disregarding the Japanese original entirely – and source new data in two ways: scrape the free preview data off an official publisher website for English translations [1] and a popular scanlation website [2]. However, this approach presents some problems on its own, as republishing the data obtained from the first source is prohibited under the terms of use of the website [3], and the quality of data in the second case is unclear [22]. What is more, none of those sources are able to provide the Japanese versions of the manga in a legal way.

[1] `https://mangaplus.shueisha.co.jp/`

[2] `https://mangadex.org/`

[3] `https://mangaplus.shueisha.co.jp/terms_web/eng/`

In an effort to legally obtain original Japanese manga scans for evaluation, we reached out to seven of the biggest Japanese publishers. All of these publishers have websites, that allow users to buy electronic versions of manga, and offer free previews of the first couple of chapters for many titles. We inquired about the possibility of using these free previews for research and redistributing them as datasets for future works, emphasizing that they would only be used for evaluation, and not for training of machine learning models. Unfortunately, all of the publishers refused to permit the use and redistribution of that data, citing reasons such as having a policy of not accepting personal requests (3 companies) or not engaging in this sort of data sharing at all (4 companies).

4.1.1 Pre-existing manga translation datasets

To date there has been only one manga translation dataset the has been made public - the OpenMantra dataset created by [40] and published on GitHub [4]. It consists of five one-shot manga volumes created by Mitsuki Kuchitaka, totalling 214 pages (1593 speech bubbles). Table 4.1 presents an overview of the titles in this dataset.

Title	Genre	#Pages	#Lines
Balloon dream	romance	38	314
Boureisougi	mystery	36	274
Rasetugari	fantasy	54	359
Tencho isoro	slice of life	40	311
Tojime no siora	battle	46	334

Table 4.1: Overview of the OpenMantra dataset [40]

In this dataset, each volume has annotations for the locations of panels and text boxes on the page, as well as the contents of the text boxes and the reading order, with professional translations into English and Chinese. We use this dataset to evaluate translation to English, splitting it into two parts - validation set (*Balloon dream* and *Tojime no siora*, 84 pages, 648 speech bubbles in total) and test set (*Boureisougi*, *Rasetugari* and *Tencho isoro*, 130 pages, 944 speech bubbles in total). This decision was made for two primary reasons: full manga volumes were necessary to evaluate whole volume translation and the settings of the stories of *Balloon dream* and *Tojime no siora* are somewhat similar to *Tencho isoro* and *Boureisougi* respectively.

4.1.2 New data

After the unsuccessful attempts of contacting the Japanese publishers, we investigated the manga datasets that do not contain translations. The biggest publicly available manga specific dataset is the Manga109-s dataset [30, 4, 74], that contains scans of 109 volumes of manga commercially published in Japan. Each volume contains annotations for, among others, the locations of panels and text boxes, as well as the contents of the text boxes. After

[4] `https://github.com/mantra-inc/open-mantra-dataset`

investigation, we found that one of the titles in the dataset – *Love Hina* – has official Polish translations for both of the volumes present in the dataset – volume 1 and volume 14 (the first and the last one in the story). After confirming with the publisher of the Polish translation that the commercial copyright for the translation has expired in Poland, we obtained second-hand physical copies of both volumes and transcribed the translations manually, as the publisher was no longer in possession of that data.

The annotations process closely followed the already existing annotations of the Japanese text. We annotated both the translated text and the reading order, as this wasn't available in the original annotations. The original lines were matched with the corresponding translated lines primarily on the basis of location, and if that was impossible, it was matched based on the content. However, sometimes the Polish edition would leave very small text untranslated, even though it was annotated in the Manga-109s dataset, in which cases we simply left the field for translation empty. The reading order was first estimated by the tool provided by [98] and then corrected by hand based on the contents of the actual speech bubbles in question.

During the annotation process, we have noticed a couple characteristics about this title and the unique challenges it would present for translation. Some of the characters in *Love Hina* speak in a dialect of Japanese, namely the Kansai dialect. According to literature, there is no consensus regarding the way of translating that dialect to Polish, with different translators choosing to express it with different dialects of Polish, often not the one that was chosen by the translator of the Polish edition of *Love Hina* [46]. Another challenge we noticed is that one of the secondary characters speaks in a manner closely resembling the way samurais would speak - a somewhat common trope in manga [24]. However, once again, there is no general consensus on how to convey that in Polish. Lastly, compared to the mangas in the OpenMantra dataset, *Love Hina* makes use of unconstrained text much more often, usually with a relatively small font. It also has more speech bubbles per page on average (7.3 in OpenMantra vs 10.5 in Love Hina, when counting only pages with text). We expect that this will have a significant impact on both the difficulty of translation and evaluation.

We have used a 50:50 validation:test split for this dataset, using the first volume as the test set and the last volume as validation set. This decision was motivated primarily by the fact that the first volume establishes the story, and so should provide a more fair benchmark for the long-context methods, as opposed to the last volume, which is severely out of context, being the 14th installment in the series. Table 4.2 presents an overview of all the data used for experiments.

Dataset	**Language**	**Set**	**#Pages**	**#Lines**
OpenMantra	English	validation	84	648
		test	130	944
Love Hina	Polish	validation	200	1895
		test	200	1810

Table 4.2: Overview of the evaluation data

4.2 Evaluation

4.2.1 Baselines

We employ four baseline methods for translation between Japanese and English. The first two baselines – `Scene-NMT` and `Scene-NMT w/ visual` – come from the original manga translation work by [40]. The first of the two methods is a transformer model that translates the contents of entire panels at once without multimodal context, while the second methods includes visual features as well. The third baseline method is the `Scene-aware NMT` proposed by [47], which also translates manga panel by panel using a transformer model, but uses the text from the previous panel as additional context. It is currently the state of the art method among NMT systems dedicated specifically to manga translation. For a more detailed description of these methods, see subsections 2.1.1 and 2.1.2. The translation outputs for all of the above methods were kindly provided by the authors of the respective works. As such we were able to use our own data splits and ensure that all the methods were evaluated equally, in a comparable way.

Following the examples of [40, 48], the last of the baseline methods we use is Google Translate (`GT`). We choose it as it supports a large range of languages, and is widely available commercially, similarly to GPT-4. It is also our only baseline for Japanese-Polish translation. All of the `GT` translations were carried out using the GoogleTranslate API with the googletrans Python library[5] and accessed in April and May 2024.

4.2.2 Automatic evaluation

For automatic evaluation we use a range of reference based metrics which we apply at sentence level. We use two lexical n-gram matching heuristic metrics – BLEU [87] and ChrF [88] – from the standardized sacreBLEU implementation [89]. Although the reliability of these two metrics has been questioned throughout the years, they are among the most frequently used metric in machine translation[40, 73, 53, 29]. Moreover, all the previous works that evaluated manga translation reported BLEU [40, 47], so we decide to follow suit, while acknowledging the limitations of this metric, such as having low correlation to human judgement and being heavily biased toward outputs using the same wording as the references. Both BLEU and ChrF give a score on the scale from 0 to 100, where higher scores correlate with higher quality of translation.

The first non-lexical machine translation evaluation metric we use is BERTScore [123], as it is considered a good representative of the embedding-based metrics category [97]. Although it is not perfect, and known to suffer for example from the antonym problem – not recognizing when the sentence has a completely different meaning from the original, because antonyms are relatively close in the vector space – it has been shown to be able to detect when the candiadte differens from reference in important content words and should be relatively well suited for scoring candidates from widely-differing systems [37]. We use the BERTScore implementation from the Python BERTScore library [6], which returns a

[5] `https://pypi.org/project/googletrans/`

[6] `https://pypi.org/project/bert-score/`

score on a $0-1$ scale, where a higher score is better.

Next, we report scores with a learned metric – BLEURT [102], specifically the top performing BLEURT-20 model [90]. The main idea behind this method, is to combine pre-training on large amounts of synthetic data with fine-tuning on human ratings. We choose this metric following the recommendation of [53]. They note that it is both a popular and a reliable choice, and complements the COMET models well as it uses a different architecure. Similarly to BERTScore, BLEURT also returns a score on a scale from 0 to 1, with results closer to 1 being better.

The last metric we report is also a learned metric – xCOMET [32]. Specifically, we use the xCOMET-XXL model [7]. xCOMET is an open-source learned metric, that in addition to standard sentence level evaluation also performs error span detection, and is currently considered the best performing publicly available metric [29]. Out of all the metrics we employ, it is the only one that calculates the score based not only on the references and the hypotheses, but also on the source text. It is also the only automatic metric other than BLEU that has been reported in a manga translation paper [47]. xCOMET returns a score on the scale from 0 to 1, where 1 means a flawless translation.

Recently, there has been a raise of LLM-based evaluators – methods, that use LLM models to score translation or text generation in general [28, 51, 76, 52]. However, due to emerging reports of those types of evaluators being heavily biased towards LLM outputs [86, 63] and difficulties with reproducing such obtained scores due to the non-public nature of the models they use [53], we forgo using those sorts of methods entirely.

4.2.3 Human evaluation

In addition to our extensive automatic evaluation, we perform human evaluation with the help of a professional Japanese-English manga translator, using the Multidimensional Quality Metrics (MQM) translation evaluation framework [11, 44].

We use a modified version of MQM-Core, tailored for manga translations with the help of researchers from Mantra Inc[8]. The error severeties remain default – Minor, Major and Cricital, but the error types were modified to suit the use case. For example, we decided to remove the Addition and Omission error types, as this is often necessary due to the constraints of the comic format. We added Formality, as formality in Japanese is very granular, and often used to express the personality of a given character (e.g. always using the formal tone, or refusing to show proper level of respect). A summary of all the error types we used can be found in Table 4.3.

The task of the human evaluator was to mark error spans on the translated lines, based on the original text and images. Each error span was assigned a category and severity. Final score was then calculated using the following MQM formula:

$$S = 1 - \frac{5 \times C_{Min.} + 10 \times C_{Maj.} + 25 \times C_{Cri.}}{\#Words} \tag{4.1}$$

[7] https://huggingface.co/Unbabel/XCOMET-XXL

[8] https://mantra.co.jp/

Error Category	Subcategories
Fluency	Punctuation Orthography Grammar
Accuracy	Mistranslation Untranslated
Proper Nouns	Orthography Failed to recognize as proper noun Terminology
Style	Formality Awkward Boring Tone
Other	Other

Table 4.3: Modified MQM error types

Where $C_{Min.}$, $C_{Maj.}$ and $C_{Cri.}$ refer to the total number of Minor, Major and Critical errors respectively, and $\#Words$ refers to the total word count in the translated text. Such calculated scores have an upper bound of 1 and no lower bound, with a higher score being preferable.

As human evaluation is very costly, we limited the evaluation to only one of the volumes in the test set – *Tencho Isoro*. For the same reason, we were only able to evaluate three sets of translations – the ground truths, the `GT` baseline and one of our methods. We decided to evaluate the ground truths to account for translator bias and to have a point of reference. We chose the `GT` baseline as the only other method available for multiple languages. As for our methods, we chose the one that was the most promising based on automatic evaluation.

Chapter 5

Results

This chapter will presents the results for all of the methods and the baselines, on translation from Japanese to English (Section 5.1) and from Japanese to Polish (Section 5.2). As requests for some of the methods are extremely costly, we report scores of one run for each of them. However, as can be seen when comparing closely related methods (e.g. PBP-VIS and PBP-VIS-NUM in Table 5.2), the methods exhibit high consistency in scores.

5.1 Japanese to English

5.1.1 Text only translation

Method	BLEU	ChrF	BERTS	BLEURT	xCOMET
GT	15.1	34.2	0.895	0.525	0.729
Scene-NMT	16.7	34.2	0.897	0.512	0.651
Scene-NMT w/ visual	16.3	34.5	0.895	0.507	0.664
Scene-aware NMT	**18.6**	**36.1**	**0.903**	0.534	0.670
LBL	13.3	32.7	0.883	0.523	0.716
PBP	15.0	36.0	0.898	**0.565**	**0.758**

Table 5.1: Results for the baselines and text-only methods

Best scores in **bold**. BERTS stands for BERTScore.

Table 5.1 presents the results of the baseline methods and the results of the text-only method proposed in this thesis. Among the methods proposed by previous work, Scene-aware NMT proposed by [47] is clearly the best method, outperforming both methods from the previous manga focused paper [40] on all of the metrics, which is consistent with the results they report. GT on the other hand, beats all three methods on xCOMET, but loses to Scene-aware NMT on all of the other metrics.

Our most basic method, LBL, performs slightly worse than GT in all aspects, but also scores higher than previous work on xCOMET. Finally, PBP shows large improvement on all of the metrics over LBL, outperforming all of the baselines on BLEURT and xCOMET, but failing to match the BLEU and BERTScore of Scene-aware NMT. It is worth noting, however, that all of the methods score considerably high on BERTScore, as even the lowest scores are already very close to the maximum score of 1.

5.1.2 Image-informed translation

Method	BLEU	ChrF	BERTS	BLEURT	xCOMET
GT	15.1	34.2	0.895	0.525	0.729
Scene-aware NMT	**18.6**	36.1	**0.903**	0.534	0.670
LBL	13.3	32.7	0.883	0.523	0.716
LBL-VIS	14.2	35.6	0.900	0.551	0.746
PBP	15.0	36.0	0.898	0.565	0.758
PBP-VIS	15.6	36.6	**0.903**	0.581	**0.776**
PBP-VIS-NUM	15.6	**36.8**	0.900	**0.582**	**0.776**

Table 5.2: Impact of the visual context

Results for the 1-shot variant. Best scores in **bold**. BERTS stands for BERTScore.

Table 5.2 compares the text-only methods and the corresponding image-informed methods. In the case of both LBL and PBP the addition of the visual context improved the scores across the board. The change, however, is much more pronounced for the more basic line by line method. Both of the image-informed methods outperform both of the baseline methods on all of the metrics, except for BLEU which is still the highest for Scene-aware NMT. It is also noticeable that in both cases, the improvements on the lexical metrics were considerably smaller, than on the learnt metrics – BLEURT and xCOMET.

PBP-VIS and PBP-VIS-NUM perform virtually the same, meaning that annotating which exact speech bubble each line came from did not yield any improvements, going against what preliminary experiments on the validation set indicated.

5.1.3 Long context translation

Method	BLEU	ChrF	BERTS	BLEURT	xCOMET
PBP-VIS	15.6	**36.6**	**0.903**	**0.581**	**0.776**
VBP-VIS-COD	15.1	35.9	0.900	0.566	0.769
VBP-VIS-3P	**16.1**	35.9	0.897	0.565	0.754
VBP-VIS-ALL	14.8	35.7	0.893	0.556	0.760
VBV-VIS	15.2	34.9	0.884	0.539	0.733

Table 5.3: Results for the long-context methods

Results for the 1-shot variant. Best scores in **bold**. BERTS stands for BERTScore.

Table 5.3 presents the results for the methods attempting to translate an entire manga volume at once and juxtaposes them with PBP-VIS for a point of reference. None of the long context methods perform better than PBP-VIS when evaluated at sentence level, the only exception being VPB-VIS-3P which performs slightly better on BLEU.

Noticeably, it seems that above the page level, more context does not necessarily mean better translation quality. VBP-VIS-COD – a method that translates the volume one page at a time, and provides a summary of the previous events as the only additional context – performs better than VBP-VIS-ALL, which gives the model access to the entire volume and all the translations produced so far. What is more, VBV-VIS, a method that translates the entire volume at once performs worse on all metrics than VBP-VIS-3P, a method that translates the volume three pages at a time. Both are also worse than the summary method.

5.1.4 Number of examples

Method	Variant	BLEU	ChrF	BERTS	BLEURT	xCOMET
LBL	1-shot	13.3	32.7	0.883	0.523	0.716
	5-shot	13.7	34.2	0.889	0.549	0.731
PBP	1-shot	15.0	36.0	0.898	0.565	0.758
	5-shot	15.6	36.6	0.898	0.565	0.752
LBL-VIS	1-shot	14.2	35.6	0.900	0.551	0.746
	5-shot	13.2	35.4	0.900	**0.581**	0.762
PBP-VIS	1-shot	15.6	36.6	**0.903**	**0.581**	0.776
	5-shot	15.5	36.5	0.902	0.580	**0.778**
PBP-VIS-NUM	1-shot	15.6	**36.8**	0.900	0.582	0.776
	5-shot	15.3	36.0	0.901	0.573	0.769
VBP-VIS-COD	1-shot	15.1	35.9	0.900	0.566	0.769
	5-shot	15.3	35.5	0.898	0.565	0.754
VBP-VIS-3P	1-shot	**16.1**	35.9	0.897	0.565	0.754
	5-shot	15.4	36.1	0.899	0.571	0.766
VBP-VIS-ALL	1-shot	14.8	35.7	0.893	0.556	0.760
	5-shot	15.6	35.6	0.888	0.559	0.755
VBV-VIS	1-shot	15.2	34.9	0.884	0.539	0.733
	5-shot	13.6	33.4	0.880	0.534	0.721

Table 5.4: Comparison of the 1-shot and 5-shot variants

Best scores in **bold**. BERTS stands for BERTScore.

Table 5.4 shows a comparison of the 1-shot and 5-shot variants of all the methods. Only the two line by line methods – LBL and LBL-VIS – show noticeable improvement when given five examples instead of one. Most of the remaining methods show only slight fluctuations of the scores, inconsistent across different metrics. The only exception is `VBV-VIS`, which shows a drop in every single score when given more examples.

5.1.5 Human evaluation

Method	Minor	Major	Critical	Score
Official	14	50	107	-1.31
`GT`	5	20	272	-4.25
`PBP-VIS`	8	18	160	-1.98

Table 5.5: Human evaluation results.

Lastly, Table 5.5 shows the results of the human evaluation for *Tencho isoro* translations. To represent the newly proposed methods, `PBP-VIS` was chosen.

The first take-away is that translation is highly subjective, as the official translations were far from a perfect score according to our judge. Keeping that in mind, `GT` had less Minor and Major errors, but almost three times as many Critical errors, and a substantially lower score. Our method fared in between the two, having a similar number of Minor and Major errors as `GT`, but 112 less Critical errors, scoring considerably better. Despite having 50% more errors than ground truths, it was considerably closer to human performance than to `GT`.

The fact that all the scores are negative and most errors are marked as Critical appears quite concerning at first, but we attribute it to a calibration issue. Since all three methods were graded by the same person, the relative differences should still be informative.

5.2 Japanese to Polish

Table 5.6 shows the results for translation to Polish. We do not report scores for `VBV-VIS` as the token limits on the completion output prevent us from using this method on the longer, 200 page volume of Love Hina. We also decided to not evaluate the `VBP-VIS-ALL` method on Polish data, as using this method is extremely costly, and none of the results on the English or Polish data indicate this method would be effective.

We notice immediately that all the scores are much lower than they were for translation to English. Having said that, we see that all of the proposed methods perform significantly better than the `GT` baseline, and that the improvement is consistent across all of the reported metrics. Once again, page by page translation (`PBP`) performs better than line by line translation (`LBL`). This is also true for the `VIS` variants of these methods.

The visual context improves the translation in the case of `LBL`, but has no impact on `PBP`, which is not consistent with the findings for English. However, once again, `PBP-VIS` and `PBP-VIS-NUM` perform virtually the same.

The long context methods perform better than the line by line methods, but worse then the best performing page by page method – `PBP-VIS`, which is exactly the same pattern as

Method	Variant	BLEU	ChrF	BERTS	BLEURT	xCOMET
`GT`	-	5.97	22.3	0.826	0.446	0.457
`LBL`	`1-shot`	7.83	24.2	0.844	0.495	0.531
	`5-shot`	7.28	24.2	0.847	0.520	0.528
`PBP`	`1-shot`	8.23	25.6	**0.852**	0.538	0.565
	`5-shot`	8.67	**25.8**	0.851	0.535	0.566
`LBL-VIS`	`1-shot`	7.28	24.9	0.845	0.515	0.543
	`5-shot`	7.67	24.6	0.846	0.517	0.540
`PBP-VIS`	`1-shot`	7.98	25.6	**0.852**	**0.539**	**0.567**
	`5-shot`	8.64	25.3	0.850	0.532	0.563
`PBP-VIS-NUM`	`1-shot`	8.25	25.7	0.851	0.532	0.566
	`5-shot`	8.59	25.7	0.850	0.532	0.564
`VBP-VIS-COD`	`1-shot`	8.28	25.1	0.846	0.523	0.550
	`5-shot`	9.01	25.6	0.844	0.510	0.555
`VBP-VIS-3P`	`1-shot`	**9.09**	25.6	0.843	0.528	0.559
	`5-shot`	8.01	25.0	0.834	0.520	0.555

Table 5.6: Results for translation to Polish

Best scores in **bold**. BERTS stands for BERTScore.

we have observed for the English data. What is more, `5-shot` does not provide a consistent improvement over `1-shot` for any of the methods.

Chapter 6

Discussion

In this chapter, we will analyze and discuss the results of the experiments, trying to explain the tendencies we notice. We will also address the Research Question, as well as the limitations of this work and the ethical considerations.

6.1 Translation to English

We will start with a high level discussion about the metrics. When looking at the results in table 5.4, it becomes quite apparent that there is little variation in the BLEU and ChrF scores, and the direction of change often does not follow that of the more widely acclaimed metrics such as BLEURT and xCOMET. We argue that n-gram based metrics might not be the best suited for evaluation of manga translation, as it often deals with short utterances, that due to the their nature of being mostly dialogue, can be translated in many ways to mean the same thing, but phrased very differently. Indeed, previous research has indicated that BLEU does not correlate with human judgement on manga data [40].

Similarly, we notice that BERTScore exhibits only very slight variation for most of the methods. However, the largest changes follow the direction of BLEURT and xCOMET. All of the methods – both the baselines and newly proposed – score very highly on BERTScore, which would indicate that it detected that the methods convey the same meaning as the reference sentences, but did not have the capabilities to meaningfully differentiate between them. This would be in line with previous reports of BERTScore having a tendency to not penalize even major mistakes, as long as there is few of them.

We will now move on to discussing the methods themselves. The first takeaway is that `PBP` performs significantly better than `LBL`. Previous research has already pointed out this might be the case, with [48] finding that translating entire paragraphs of literary text at once results in higher quality translation than sentence by sentence translation. We theorised it might not be the case for `LBL` and `PBP` as we included the manga context in the prompt for both, but results prove that the findings of [48] transfer to the manga domain as well.

The second takeaway is that the inclusion of visual context does significantly improve translation quality. Interestingly, the gain from showing the image to the model is smaller than the gain from just giving it more text as context. However, combining both achieves the best results. [40] argued that manga translation systems would benefit from visual context, but their own method did not incorporate a strong enough visual model to achieve that. Our findings show that the visual capabilities of GPT-4-Turbo are sufficiently good to effectively leverage that extra information.

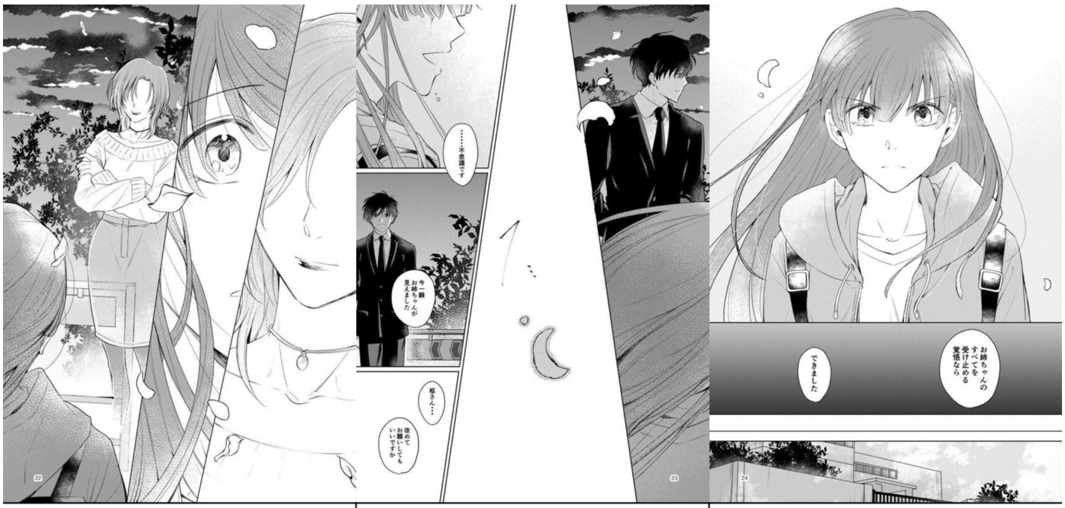
As mentioned in the previous chapter, `PBP-VIS` and `PBP-VIS-NUM` performing exactly the same was unexpected after our experiments on the validation set. We argue that as there was no drop in quality, the model was not making any use of the text on the page. As for why there was no improvement, we hypothesize that only specific types of manga benefit from this feature under the setup that was tested. A larger corpus would be necessary to prove that. However, the most recent GPT model was released midway through our experiments, so it is also possible that the improvement we saw during experiments was an effect that is limited to the previous model – `gpt-4-0125-preview`.

Continuing the trend, the long context methods were consistently worse than the method, they were building upon – `PBP-VIS`. When designing the long context methods, we expected two types of benefits. Firstly, we expected that maintaining a broader overview of the work would result in the model being more consistent in narration across the entire volume. However, this is something that only humans can realistically measure, and our resources were too limited to adequately evaluate that.

Secondly, we expected that empowering the model with volume-level context would allow it to make more informed translation decisions on sentence level as well. To a certain level, we were able to achieve that – an example can be seen on Fig. 6.1. `PBP-VIS` is completely confused, as to who the sister in question is and whose sister it is. As the only context it has is an image of a woman, it makes an understandable assumption that the woman is the subject of the sentence. However, `VBP-VIS-COD` makes use of the additional context of the summary it has accumulated thus far:

> In a ramen-centric world, Tsubame challenges funeral norms revealing societal strains, while Kururu debunks ghost myths amidst a deadly sibling dispute. Concurrently, an aspiring makeup artist, blaming her ambitions for **her sister's death**, is consoled by Tsubame, affirming kinship. Amid grief, Kururu sees a fleeting **vision of her sister**, hears mysterious whispers, and resolves to seek further understanding.

and correctly resolves that the ”sister” is not the speaker herself, but her relative. However, the same example also illustrates a possible reason for why `VBP-VIS-COD` performs worse than `PBP-VIS` on the whole – the summaries are not always correct. In this particular case, the names are mixed up. *Kururu* is the name of the man, not the woman, and *Tsubame* is the name of the late sister, not the man. Additionally, the phrase *ramen-centric world*, stemming from the early events of the story taking place at a ramen store, is a pretty vague and inaccurate description. These sort of misunderstandings will be carried on to subsequent summaries and might negatively impact translations down the line. We hypothesize that the setup of `VBP-VIS-COD` makes mistakes much more costly, as they are carried over as context for future translations, which ends up outweighing the potential benefits from being able to resolve ambiguities. For instance, we noticed that the the name of the author consistently


……不思議です
今一瞬お姉ちゃんが見えました
枢さん…
改めてお願いしてもいいですか
お姉ちゃんのすべてを受け止める覚悟なら
できました

REFERENCE	“...It's strange.” “i just saw my sister” “Kururu-san” “can i ask a favor of you again?”	“I'm ready to accept everything **about my sister**” “i'm ready for it”
PBP-VIS	“It's mysterious...” “i just saw my sister for a moment.” “Mr. Kuru...” “May I ask you again?”	“If you're prepared to accept everything **about me as your older sister**,” “I am ready.”
VBP-VIS-COD	“...It's mysterious.” “I just saw my sister for a moment.” “Kururu…” “May I ask you again?”	“If I'm prepared to accept everything **about my sister**” “I am ready”

Figure 6.1: A manga page with subject elipsis.

©Mitsuki Kuchitaka, from the OpenMantra dataset [40].

appeared in summaries, attributed to one of the characters, due to the model misinterpreting the title page for a character introduction.

However, this logic would not apply to `VBP-VIS-3P` and `VBV-VIS`, as none of those methods are given any context produced by the model itself. And yet, it appears as though the more pages are given to the model, the worse it performs, with `VBV-VIS` performing worse than `LBL-VIS`. We attribute that to two possible reasons. First – a property of GPT-4 that was reported by [116] – they find that the model can lose focus for multiple images, which we notice in the scores, as the performance of both `VBP-VIS-ALL` and `VBP-VIS-3P` is reduced to the level comparable to `PBP`. The further performance drop of `VBV-VIS` compared to `VBP-VIS-ALL` is likely due to the lack of guidance in the prompt and examples, as we instruct all of the other `VIS` methods to ground their translation in the image context, by explaining how it informs the translation. We were not able to do that for `VBV-VIS` due to the token limits on the completion, which is exactly what `VBP-VIS-ALL` was designed to make up for. The second reason might be that GPT-4 was simply not trained for extremely long completions – a category that both `VBP-VIS-3P` and `VBV-VIS` fall into, as we were often hitting the limit in our experiments, necessitating retries.

Next, we will discuss the effect of adding more examples in the prompt. The only methods that improved meaningfully and consistently were the line by line methods. The work of [39] reports a similar finding and hypothesize that giving the model more text to work with makes additional examples redundant, as enough context is provide by the long text itself. Our findings show this is also true for manga translation. Interestingly, `LBL-VIS` performed almost on par with `PBP-VIS` when given 5 examples, which proves how crucial additional context is. However, neither `PBP-VIS`, nor any of the long context methods showed a similar improvement, which once again shows that there is a point of context ”saturation”, after which there are diminishing returns, or even a decrease in quality.

Lastly, we will address the human evaluation. The results of automatic evaluation led us to believe that `PBP-VIS` would perform considerably better than `GT`, but the difference is much more pronounced when judged by a human translator. After manually inspecting the error annotations, we see that most errors made by `GT` are Mistranslations (147), with Orthography (84) and Awkward (52) coming in second and third. In the case of `PBP-VIS`, the order was: Awkward (77), Mistranslation (44) and Orthography (40). This leads us to believe, that our method is much better in getting the meaning across, and struggles mostly with style, while `GT` is often completely wrong. Coincidentally, the top 3 error categories for the ground truths were: Awkward (84), Mistranslation (43) and Other (15), proving just how subjective translation is. It also raises a question of how big the impact of the quality of the references was on the automatic evaluation. However, answering that would require more extensive human evaluation.

6.2 Translation to Polish

The fact that the scores are lower on average was expected to a certain degree as English is the most represented language in the training data and Polish is a more niche and challenging language [106, 48]. Indeed, even the baseline method scores significantly lower than it did for English.

When comparing the newly proposed methods against each other, we see that they all score much more similar compared to how they did for English. Most noticeably, `PBP` and `PBP-VIS` score virtually the same. We hypothesize there are three main reasons for this.

First of all, the ability of GPT-4 to model the Polish language imposes a limit on the quality of translation we can obtain, and the lack of further improvement might not be due to factual errors, but linguistic errors. After inspecting the translations manually, we noticed that the translations often sound awkward, or appear to be calques from English or Japanese, which the reader would understand, but are not grammatically correct.

Second of all, the ground truth translations for the polish data are 17 years old at the time of testing, and follow a particular translation style that might not be in line with today’s language. We argue it is possible that the translations score lower because they follow a very different style, which the metrics view as ”wrong” when compared to ground truths. For example, when discussing the dataset we mentioned that some of the characters speak in the Kansai dialect, that the Polish translator decided to localise as a highlander dialect, but we notice the model simply uses standard Polish for everything.

Lastly, the author of *Love Hina* makes use of unconstrained text more often, and on average puts more text on the page than the author of the mangas in OpenMantra. Most often, this is because there are more characters at once on a single page of *Love Hina*. This makes the overall text more convoluted, as each of the speakers usually follows a separate chain of thought, making it more challenging to translate. We tried to tackle it by asking the model to specify details about each of the speakers, such as gender and name, but the model was reluctant to follow that format.

6.3 Research Question

Having summed up the findings, we can now answer to Research Question that was posed at the beginning of this thesis paper in Section 1.1. To reiterate, the main and the sub-questions were as follows:

How effective are multimodal LLMs in translation scenarios where incorporating multimodal context is necessary to bridge the information gap between typologically distant languages?

1. How big is the impact of visual context on the quality of translation produced by state of the art multimodal large language models?

2. How well can such models scale up to translating long narratives?

3. How consistent is the performance across different language pairs?

The answer to the first sub-question is: substantial and positive, but up to a point. We notice that the methods benefited from including the visual context over the no-image baselines, but there was no point in providing more than one page.

The answer to the second sub-question is: they can translate them on page by page basis, but as far as we have been able to establish, trying to introduce consistency across multiple pages leads to a worse translation.

The answer to the last sub-question is: inconsistent and limited by the capabilities of modelling the given language. We noticed that the results from the experiments on the English data only transferred to a limited extent to the Polish data, and the overall quality was worse. That being said, it still vastly outperformed the baseline and set a new standard.

Finally, to answer the main question – the investigated model - GPT-4-Turbo - is effective on a page by page basis, but for the time being lacks the ability to meaningfully process multi-page narratives. Furthermore, the improvement is much less pronounced for a niche language such as Polish, due to the model's limited ability to use it correctly and naturally overshadowing the potential benefits.

6.4 Limitations

We will now discuss the limitations of this study, as well as the limitations of the proposed methods.

The first limitation of this study is the amount of data we used for tasting. Although we make meaningful contributions to solve this problem, there is still a severe lack of evaluation data, and as such it is hard to say how consistent the effects we find would be across the works of different authors and genres. What is more, we only investigate one language other than English, limited by our ability to manually inspect the output and analyse the models mistakes for other languages. Another limiting factor is that reference based evaluation is limited by the quality of the references, and in the case of literary translation, there are often multiple, very different but equally correct ways of translating a given sentence . As such, it is possible that at times the models outputs were overly penalized [29]. To a certain extent, we tried to address it by carrying out human evaluation, but its scope was very limited. Lastly, we acknowledge that learned metrics might exhibit unknown biases, depending on the data they were trained on.

In terms of the limitations of the proposed methods, LLM based systems are known to have several problems stemming from how much compute is necessary to run these models. These include high energy consumption, environmental impact and monetary costs [19, 105, 8], as well as the latency that API calls introduce in those systems. Some of the methods we tested took over two minutes on average to process one page, meaning that as they are now, they are not suitable for an online application. Lastly, the reliability of these systems, both in terms of the API availability and the quality of the outputs are an unsure matter, as they rely completely on the service provided by the company that hosts the models.

6.5 Ethical Considerations

The ethics of AI have become an important matter to consider when developing those technologies, and are as relevant as ever when it comes to large language models [8, 9]

Large language models are generally known to suffer from issues such as biased outputs, toxic statements and hallucinations [9, 17, 109, 31, 7]. These are problematic in all applications, but in the case of literary translation there is an additional stakeholder – the author of the original text – that stands to suffer from being misrepresented, hurting the perception of their work and values [107].

The next consideration is the impact of this technology on the role of human translators. While we, as well as some of the previous works in this field ([48]), do not recommend LLM translation systems as a substitute for the human translator, but as assistance to speed up the process, we are aware that it might threaten their role and negatively impact the job market, instead of making the works more readily accessible.

Lastly, as the training data for GPT-4 was never fully disclosed, it is not out of question that copyrighted content was a part of it, and whether or not that would be legal and considered fair-use is a subject of an ongoing debate [111, 82]. We recognize this issue, especially relevant in the the field of manga translation, where almost all of the works are heavily protected by copyright laws.

Chapter 7

Conclusion

This chapter gives an overview of the project's contributions. Next, we will reflect on the results and draw some conclusions. Finally, promising directions for future work will be discussed.

7.1 Contributions

To the best of our knowledge, this work is the first to evaluate the capabilities of GPT-4-Turbo for multimodal machine translation. We have made the following contributions in this field:

- **State of the art multimodal multilingual manga machine translation system**

  Our best methods vastly outperform previously existing SOTA baselines on all metrics we tested. Moreover, it is the first method for manga to support languages other than English.

- **First publicly available manga translation evaluation suite**

  The code artifacts we contribute can be used both to benchmark the possibilities of other LLMs for manga, as well as compare them against other methods.

- **The first ever parallel Japanese-Polish manga translation evaluation dataset**

  The annotations we contribute, when coupled with the Manga-109s dataset, contain over 2 times more lines than the only previously existing publicly available dataset with professional translations - OpenMantra.

7.2 Conclusions

Our findings show that GPT-4-Turbo is able to produce comparably high quality translations to English, and is better than any existing alternatives for both of the languages we evaluated - English and Polish. Moreover, we have been able to prove that it is able to successfully improve the translation by incorporating the visual modality.

After manually inspecting the outputs, we consider it to be the ”better than alternatives” in terms of automated manga translation. Having said, it is by no means a substitute for a professional human translator as it tends to critically misunderstand parts of the narrative, making it unsuitable for translating long narratives without supervision. We think it would be best used as a basis for a language learning tool, or as an aid to a professional human translator. It also makes for a SOTA baseline for future work.

7.3 Future work

We will now list and discuss promising areas for future work, that we have come across in our research:

Other LLMs – The first, and perhaps most obvious direction is evaluating other multimodal large language models. We have done some limited experimenting with Gemini 1.5 Pro [92] and consider it very promising for this use case. The recently released GPT-4o should also be evaluated, as it could help bring down the cost of operation [1].

More manga data – Focused efforts into finding a way to create more parallel manga data are necessary. Possibles avenues include annotating translations of *Love Hina* into other languages as it is a popular title, available in many languages. However, this could be one for other mangas from the Manga-109s [30, 4, 74] as well. We know of at least two more titles in that dataset that have professional translations to other languages - *Highschool! Kimengumi* to French[1] and *Salad Days* to Indonesian[2]. However, a more large scale collaboration with Japanese publishers would probably be a more scalable solution.

Better evaluation frameworks – Manga translation is a very unique use case, and as such we fear that both automatic methods, as well as frameworks for involving human experts might not address this issue well enough and warrant dedicated solutions.

Translation Memory with fuzzy matching – Previous works have indicated that using fuzzy matching to select examples to show the model from Translation Memory has produced promising results [77, 66]. In this work, we have chosen to focus on highly complicated, handcrafted examples (e.g. creating example explanations for the `VIS` methods) due to the small amount of data and the large diversity of it, but we think TM with fuzzy matching could work well in the manga use case as well.

Interactive MT – Interactive machine translation is another avenue that we have not investigated in this work, but has been proposed by other works for non-manga use cases [66]. One example scenario, that we think could work well, is having a human expert translate the first couple of pages of a given manga and letting the model work with that as a guide, asking for instruction in critical moments of the translation.

[1] https://fr.wikipedia.org/wiki/Kimengumi

[2] https://en.wikipedia.org/wiki/Salad_Days_(manga)

Bibliography

[1] Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

[2] マンガに特化した多言語翻訳システム『mantra engine』が小学館『マンガワン』の英語版展開を支援. https://prtimes.jp/main/html/rd/p/000000004.000059295.html. Accessed: 2024-04-29.

[3] Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. From sparse to dense: Gpt-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269*, 2023.

[4] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020. doi: 10.1109/mmul.2020.2987895.

[5] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[6] Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. Doubly attentive transformer machine translation. *arXiv preprint arXiv:1807.11605*, 2018.

[7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[9] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[10] Atsuko Suga Borgmann. *Multimodal Vocabulary Learning Through Manga in Japanese as a World Language*. PhD thesis, The University of Wisconsin-Milwaukee, 2023.

[11] Aljoscha Burchardt. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL `https://aclanthology.org/2013.tc-1.6`.

[12] Iacer Calixto, Desmond Elliott, and Stella Frank. Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, 2016.

[13] Iacer Calixto, Qun Liu, and Nick Campbell. Incorporating global visual features into attention-based neural machine translation. *arXiv preprint arXiv:1701.06521*, 2017.

[14] Jiali Chen, Ryo Iwasaki, Naoki Mori, Makoto Okada, and Miki Ueno. Understanding multilingual four-scene comics with deep learning methods. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 32–37. IEEE, 2019.

[15] Wei-Ta Chu and Wei-Wei Li. Manga facenet: Face detection in manga based on deep neural network. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 412–415, 2017.

[16] Wei-Ta Chu and Chih-Chi Yu. Text detection in manga by deep region proposal, classification, and regression. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2018.

[17] Marta R Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. Toxicity in multilingual machine translation at scale. *arXiv preprint arXiv:2210.03070*, 2022.

[18] Hadi Akbar Dahlan. The publishing and distribution system of japanese manga and doujinshi. *Publishing Research Quarterly*, 38(4):653–664, 2022.

[19] Alex de Vries. The growing energy footprint of artificial intelligence. *Joule*, 7(10): 2191–2194, 2023.

[20] Julián Del Gobbo and Rosana Matuk Herrera. Unconstrained text detection in manga: a new dataset and baseline. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 629–646. Springer, 2020.

[21] Jean-Benoit Delbrouck and Stéphane Dupont. Umons submission for wmt18 multimodal translation task. *arXiv preprint arXiv:1810.06233*, 2018.

[22] Jeremy Douglass, William Huber, and Lev Manovich. Understanding scanlation: How to read one million fan-translated manga pages. *Image & Narrative*, 12(1): 190–227, 2011.

[23] David Dubray and Jochen Laubrock. Deep cnn-based speech balloon detection and segmentation for comic books. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1237–1243. IEEE, 2019.

[24] Patrycja Duc-Harada. Znaczenie i wpływ jezyka postaci (yakuwarigo) na kształtowanie kompetencji jezykowych studentów japonistyki w polsce. *Ogrody Nauk i Sztuk*, (9):301–319, 2019.

[25] Arpita Dutta, Samit Biswas, and Amit Kumar Das. Cnn-based segmentation of speech balloons and narrative text boxes from comic book page images. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(1):49–62, 2021.

[26] Paweł Dybała. The translator is wrong!: readers' attitudes towards official manga translations biased by fan-made scanlations. *Relacje Miedzykulturowe= Intercultural Relations*, 4(2 (8)), 2020.

[27] D Elliott, S Frank, and E Hasler. Multi-language image description with neural sequence models. corr. *arXiv preprint arXiv:1510.04709*, 2015.

[28] Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.100. URL https://aclanthology.org/2023.wmt-1.100.

[29] Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, 2023.

[30] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Manga109 dataset and creation of metadata. In *Proceedings of the 1st international workshop on comics analysis, processing and understanding*, pages 1–5, 2016.

[31] Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023.

[32] Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*, 2023.

[33] Hongcheng Guo, Boyang Wang, Jiaqi Bai, Jiaheng Liu, Jian Yang, and Zhoujun Li. M2c: Towards automatic multimodal manga complement. *arXiv preprint arXiv:2310.17130*, 2023.

[34] Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. Bridging the gap between synthetic and authentic images for multimodal machine translation. *arXiv preprint arXiv:2310.13361*, 2023.

[35] Jeremy Gwinnup and Kevin Duh. A survey of vision-language pre-training from the lens of multimodal machine translation. *arXiv preprint arXiv:2306.07198*, 2023.

[36] Jeremy Gwinnup, Tim Anderson, Brian Ore, Eric Hansen, and Kevin Duh. Enhancing video translation context with object labels. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 130–137, 2023.

[37] Michael Hanna and Ondřej Bojar. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, 2021.

[38] Jindřich Helcl, Jindřich Libovickỳ, and Dušan Variš. Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*, 2018.

[39] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.

[40] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards fully automated manga translation, 2021.

[41] Xin Huang, Jiajun Zhang, and Chengqing Zong. Contrastive adversarial training for multi-modal machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[42] Shih-Hsuan Hung, Yu-Chi Lai, Shih-Chang Wong, Chia-Hsing Chiu, and Chih-Yuan Yao. Arbitrary screen-aware manga reading framework with parameter-optimized panel extraction. *IEEE MultiMedia*, 26(2):55–65, 2019.

[43] Hikaru Ikuta, Runtian Yu, Yusuke Matsui, and Kiyoharu Aizawa. Towards content-aware pixel-wise comic panel segmentation. In *International Conference on Pattern Recognition*, pages 7–21. Springer, 2022.

[44] ISO 5060:2024. Translation services — Evaluation of translation output — General guidance. Standard, International Organization for Standardization, Geneva, CH, February 2024.

[45] Julia Ive, Pranava Madhyastha, and Lucia Specia. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*, 2019.

[46] Hanna JAŚKIEWICZ. Reprezentacja dialektu bawarskiego i dialektu kansai w literaturze współczesnej w kontekście ideologii jezykowych w niemczech i japonii. In *Forum Filologiczne ATENEUM*, 2020.

[47] Hiroto Kaino, Soichiro Sugihara, Tomoyuki Kajiwara, Takashi Ninomiya, Joshua B Tanner, and Shonosuke Ishiwatari. Utilizing longer context than speech bubbles in automated manga translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17337–17342, 2024.

[48] Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*, 2023.

[49] Jin Kato, Motoi Iwata, and Koichi Kise. Manga vocabulometer, a new support system for extensive reading with japanese manga translated into english. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, pages 223–235. Springer, 2021.

[50] U-Ram Ko and Hwan-Gue Cho. Sickzil-machine: a deep learning based script text isolation system for comics translation. In *Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14*, pages 413–425. Springer, 2020.

[51] Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL `https://aclanthology.org/2023.wmt-1.64`.

[52] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.

[53] Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. *arXiv preprint arXiv:2401.06760*, 2024.

[54] Samu Kovanen and Kiyoharu Aizawa. A layered method for determining manga text bubble reading order. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4283–4287. IEEE, 2015.

[55] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. On vision features in multimodal machine translation. *arXiv preprint arXiv:2203.09173*, 2022.

[56] Lin Li, Turghun Tayir, Yifeng Han, Xiaohui Tao, and Juan D Velasquez. Multimodality information fusion for automated machine translation. *Information Fusion*, 91: 352–363, 2023.

[57] Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. Visa: An ambiguous subtitles dataset for visual scene-aware machine translation. *arXiv preprint arXiv:2201.08054*, 2022.

[58] Yihang Li, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi, and Wei Li. Video-helpful multimodal machine translation. *arXiv preprint arXiv:2310.20201*, 2023.

[59] Brian Lim. Multimodal models for learning to order sentences in manga.

[60] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[61] Jiayun Liu. Diseño y desarrollo de un traductor de comics. Unpublished, June 2022. URL `https://oa.upm.es/71255/`.

[62] Xueting Liu, Chengze Li, Haichao Zhu, Tien-Tsin Wong, and Xuemiao Xu. Text-aware balloon extraction from manga. *The Visual Computer*, 32(4):501–511, 2016.

[63] Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. Llms as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*, 2023.

[64] Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*, 2022.

[65] Chenyang Lyu, Jitao Xu, and Longyue Wang. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*, 2023.

[66] Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, Siyou Liu, and Longyue Wang. A paradigm shift: The future of machine translation lies with large language models, 2024.

[67] Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. Improving end-to-end text image translation from the auxiliary text translation task. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1664–1670. IEEE, 2022.

[68] Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. Modal contrastive learning based end-to-end text image machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[69] Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. E2timt: Efficient and effective modal adapter for text image machine translation. In *International Conference on Document Analysis and Recognition*, pages 70–88. Springer, 2023.

[70] Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *International Conference on Document Analysis and Recognition*, pages 484–501. Springer, 2023.

[71] Anna Marchionda. Crowdsourcing and relay translation of manga: An analysis the special one-shot death note never complete.

[72] Gema Marín Núñez. Desarrollo de un sistema de traducción automática de manga mediante algoritmos de deep learning y técnicas ocr. 2021.

[73] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*, 2020.

[74] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. doi: 10.1007/s11042-016-4020-z.

[75] Loitongbam Sanayai Meetei, Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. Do cues in a video help in handling rare words in a machine translation system under a low-resource setting? *Natural Language Processing Journal*, 3: 100016, 2023.

[76] John Mendonça, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. Simple llm prompting is state-of-the-art for robust and multilingual dialogue evaluation. *arXiv preprint arXiv:2308.16797*, 2023.

[77] Yasmin Moslem, Rejwanul Haque, and Andy Way. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*, 2023.

[78] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Comic mtl: optimized multi-task learning for comic book image analysis. *International Journal on Document Analysis and Recognition (IJDAR)*, 22:265–284, 2019.

[79] Andrei Novikov. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230, apr 2019. doi: 10.21105/joss.01230. URL `https://doi.org/10.21105/joss.01230`.

[80] Toru Ogawa, Atsushi Otsubo, Rei Narita, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object detection for comics using manga109 annotations. *arXiv preprint arXiv:1803.08670*, 2018.

[81] Naoki Ono, Kiyoharu Aizawa, and Yusuke Matsui. Comic image inpainting via distance transform. In *SIGGRAPH Asia 2021 Technical Communications*, pages 1–4. 2021.

[82] David W Opderbeck. Copyright in ai training data: A human-centered approach. *Oklahoma Law Review*, 76, 2024.

[83] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Rad-

ford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[84] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[85] Xufang Pang, Ying Cao, Rynson WH Lau, and Antoni B Chan. A robust panel extraction method for manga. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1125–1128, 2014.

[86] Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

[87] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[88] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

[89] Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

[90] Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for mt. *arXiv preprint arXiv:2110.06341*, 2021.

[91] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

[92] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[93] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[94] Christophe Rigaud, Jean-Christophe Burie, and Jean-Marc Ogier. Text-independent speech balloon segmentation for comics and manga. In *Graphic Recognition. Current Trends and Challenges: 11th International Workshop, GREC 2015, Nancy, France, August 22–23, 2015, Revised Selected Papers 11*, pages 133–147. Springer, 2017.

[95] Christophe Rigaud, Nhu-Van Nguyen, and Jean-Christophe Burie. Text block segmentation in comic speech bubbles. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, pages 250–261. Springer, 2021.

[96] Christian Roggia and Fabio Persia. Extraction of frame sequences in the manga context. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 96–99. IEEE, 2020.

[97] Hadeel Saadany and Constantin Orasan. Bleu, meteor, bertscore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. *arXiv preprint arXiv:2109.14250*, 2021.

[98] Ragav Sachdeva and Andrew Zisserman. The manga whisperer: Automatically generating transcriptions for comics. *arXiv preprint arXiv:2401.10224*, 2024.

[99] Masaki Saito and Yusuke Matsui. Illustration2vec: a semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015.

[100] Júlia Sato, Helena Caseli, and Lucia Specia. Choosing what to mask: More informed masking for multimodal machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 244–253, 2023.

[101] Simone Schroff. An alternative universe? authors as copyright owners-the case of the japanese manga industry. *Creative Industries Journal*, 12(1):125–150, 2019.

[102] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

[103] Mhd Saeed Sharif, Bilyaminu Auwal Romo, Harry Maltby, and Ali Al-Bayatti. An effective hybrid approach based on machine learning techniques for auto-translation: Japanese to english. In *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 557–562. IEEE, 2021.

[104] Conghao Tom Shen, Violet Yao, and Yixin Liu. Maru: A manga retrieval and understanding system connecting vision and language. *arXiv preprint arXiv:2311.02083*, 2023.

[105] Jovan Stojkovic, Esha Choukse, Chaojie Zhang, Inigo Goiri, and Josep Torrellas. Towards greener llms: Bringing energy-efficiency to the forefront of llm inference. *arXiv preprint arXiv:2403.20306*, 2024.

[106] Patrycja Świeczkowska. Towards a direct japanese-polish machine translation system. In *Proceedings of the 8th Language & Technology Conference*, 2017.

[107] Kristiina Taivalkoski-Shilov. Ethical issues regarding machine (-assisted) translation of literary texts. *Perspectives*, 27(5):689–703, 2019.

[108] Takamasa Tanaka, Kenji Shoji, Fubito Toyama, and Juichi Miyamichi. Layout analysis of tree-structured scene frames in comic images. In *IJCAI*, volume 7, pages 2885–2890. Citeseer, 2007.

[109] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*, 2024.

[110] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. *arXiv preprint arXiv:1708.05943*, 2017.

[111] Andrew W Torrance and Bill Tomlinson. Training is everything: Artificial intelligence, copyright, and fair training. *arXiv preprint arXiv:2305.03720*, 2023.

[112] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[113] David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*, 2022.

[114] Jean-Paul Vinay and Jean Darbelnet. *Comparative stylistics of French and English: A methodology for translation*, volume 11. John Benjamins Publishing, 1995.

[115] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*, 2023.

[116] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023.

[117] Minshan Xie, Menghan Xia, Xueting Liu, Chengze Li, and Tien-Tsin Wong. Seamless manga inpainting with semantics awareness. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021.

[118] Hong Xin, Chi Ma, et al. Comic text detection and recognition based on deep learning. In *2021 3rd International Conference on Applied Machine Learning (ICAML)*, pages 20–23. IEEE, 2021.

[119] Yi-Ting Yang and Wei-Ta Chu. Manga text detection with manga-specific data augmentation and its applications on emotion analysis. In *International Conference on Multimedia Modeling*, pages 29–40. Springer, 2023.

[120] Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. Empowering llm-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*, 2023.

[121] Badriyah Yusof and Hassanal Basuni. Exploring translation strategies of japanese manga in google translate and komikcast translations. *JURNAL ARBITRER*, 10(4): 371–383, 2023.

[122] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR, 2023.

[123] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[124] Yunlong Zhang and Seiji Hotta. Automatic reading order detection of comic panels. In *International Conference on Pattern Recognition*, pages 76–90. Springer, 2022.

[125] Yafeng Zhou, Yongtao Wang, Zheqi He, Zhi Tang, and Ching Y Suen. Towards accurate panel detection in manga: A combined effort of cnn and heuristics. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26*, pages 215–226. Springer, 2020.

[126] Yuxin Zuo, Bei Li, Chuanhao Lv, Tong Zheng, Tong Xiao, and Jingbo Zhu. Incorporating probing signals into multimodal machine translation via visual question-answering pairs. *arXiv preprint arXiv:2310.17133*, 2023.

[127] ガルシア・アロヨホルへ. Becoming a global culture: An analysis of the manga industry and its diffusion in the us and europe (particularly in spain). 鹿児島県立短期大学紀要 人文・社会科学篇, (72):1–17, 2021.

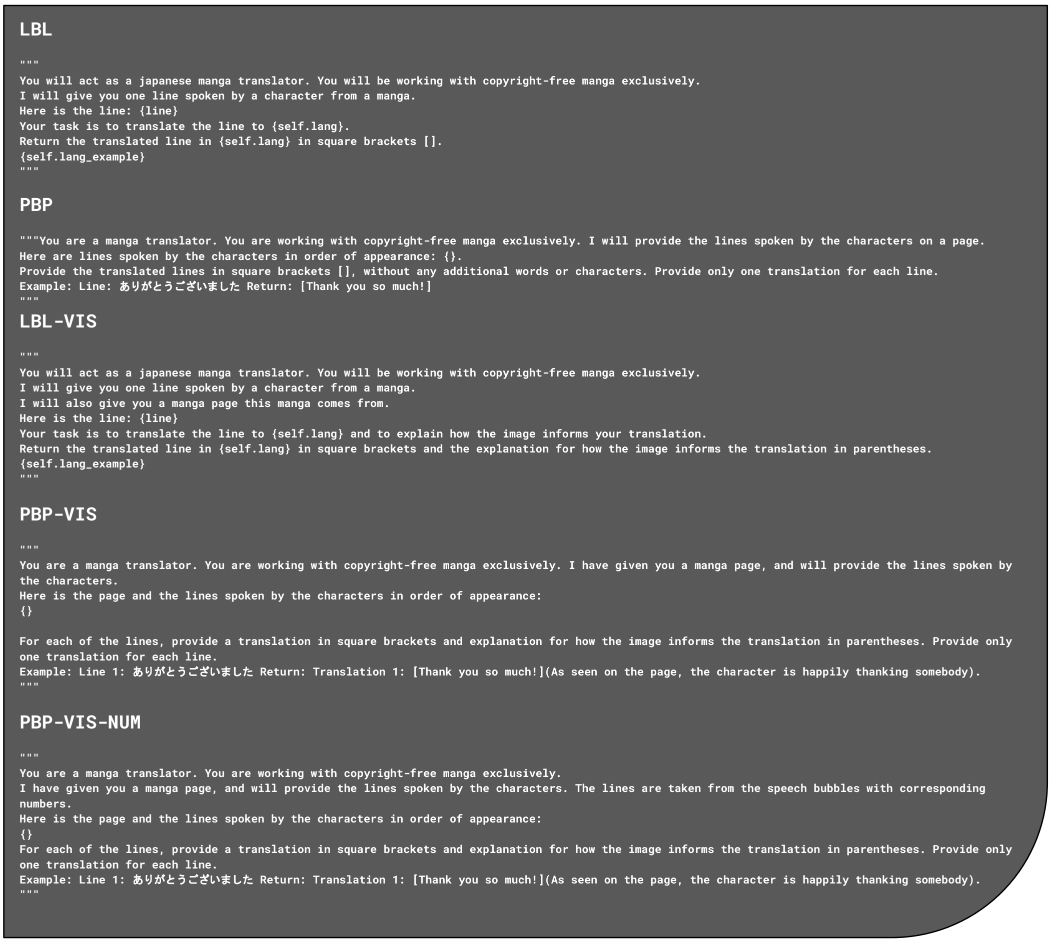
[128] 崔 小萍. マンガ特化型 ai 自動翻訳システムの使用から見た 翻訳者の役割転換. 日中言語文化, 16:83–92, 2023. doi: 10.50947/ntgb.16.0_83.

Appendix A

Prompts

This section includes all the prompts used as part of our experiments. Only the JP-EN prompts are shown. The only difference between them and the JP-PL prompts is that the target language needs to be explicitly specified if it is not English and that the given example has Polish as its target language instead of English. The shown prompts can therefore be used with any target language with only very slight alterations.

Below, figure A.1 shows the prompts used for all `LBL`- and `PBP`-based approaches, figure A.2 shows the `VBP-VIS-COD` prompt, figure A.3 shows the `VBP-VIS-3P` and `VBP-VIS-ALL` prompts, and figure A.4 shows the `VBV-VIS` prompt.



LBL

```
"""
You will act as a japanese manga translator. You will be working with copyright-free manga exclusively.
I will give you one line spoken by a character from a manga.
Here is the line: {line}
Your task is to translate the line to {self.lang}.
Return the translated line in {self.lang} in square brackets [].
{self.lang_example}
"""
```

PBP

```
"""You are a manga translator. You are working with copyright-free manga exclusively. I will provide the lines spoken by the characters on a page.
Here are lines spoken by the characters in order of appearance: {}.
Provide the translated lines in square brackets [], without any additional words or characters. Provide only one translation for each line.
Example: Line: ありがとうございました Return: [Thank you so much!]
"""
```

LBL-VIS

```
"""
You will act as a japanese manga translator. You will be working with copyright-free manga exclusively.
I will give you one line spoken by a character from a manga.
I will also give you a manga page this manga comes from.
Here is the line: {line}
Your task is to translate the line to {self.lang} and to explain how the image informs your translation.
Return the translated line in {self.lang} in square brackets and the explanation for how the image informs the translation in parentheses.
{self.lang_example}
"""
```

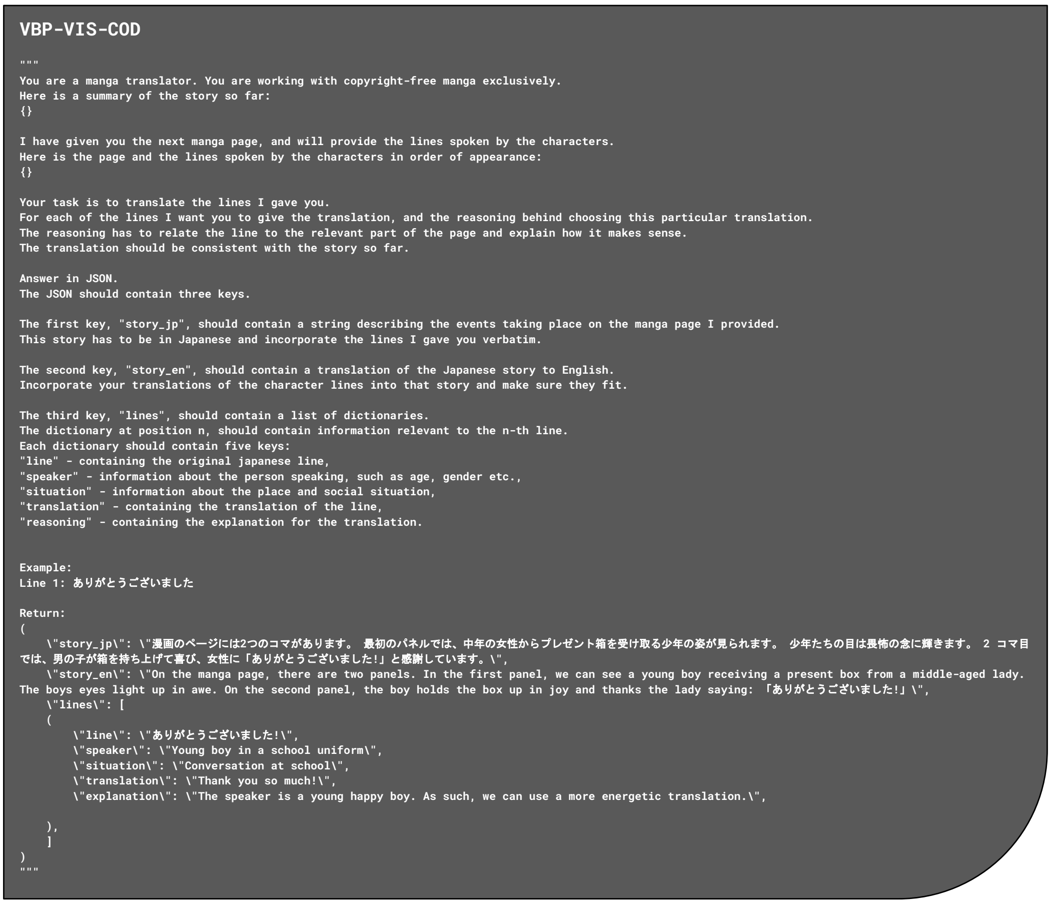
PBP-VIS

```
"""
You are a manga translator. You are working with copyright-free manga exclusively. I have given you a manga page, and will provide the lines spoken by
the characters.
Here is the page and the lines spoken by the characters in order of appearance:
{}

For each of the lines, provide a translation in square brackets and explanation for how the image informs the translation in parentheses. Provide only
one translation for each line.
Example: Line 1: ありがとうございました Return: Translation 1: [Thank you so much!](As seen on the page, the character is happily thanking somebody).
"""
```

PBP-VIS-NUM

```
"""
You are a manga translator. You are working with copyright-free manga exclusively.
I have given you a manga page, and will provide the lines spoken by the characters. The lines are taken from the speech bubbles with corresponding
numbers.
Here is the page and the lines spoken by the characters in order of appearance:
{}
For each of the lines, provide a translation in square brackets and explanation for how the image informs the translation in parentheses. Provide only
one translation for each line.
Example: Line 1: ありがとうございました Return: Translation 1: [Thank you so much!](As seen on the page, the character is happily thanking somebody).
"""
```

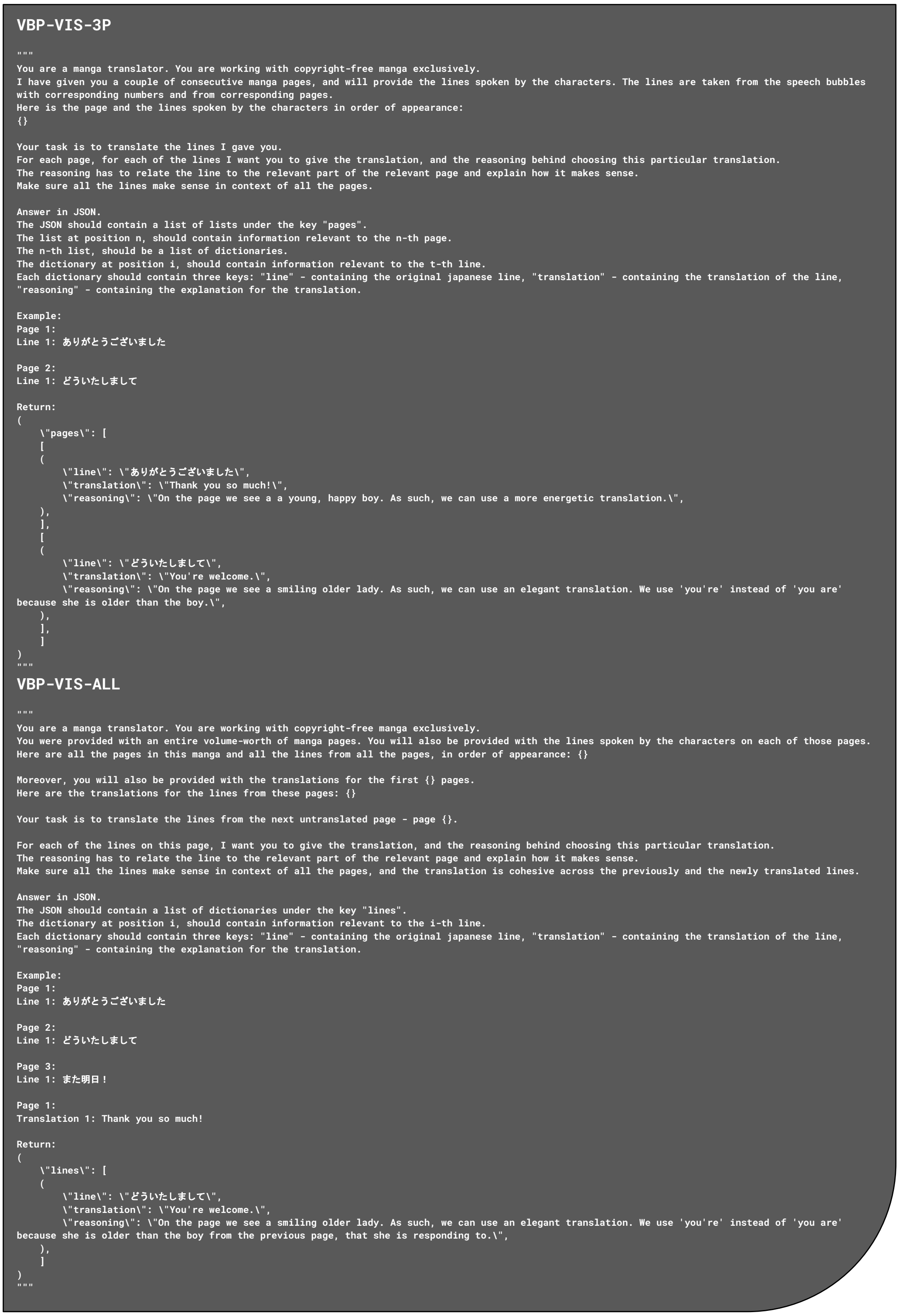
Figure A.1: Prompts used for all LBL- and PBP-based approaches.



VBP-VIS-COD

```
"""
You are a manga translator. You are working with copyright-free manga exclusively.
Here is a summary of the story so far:
{}

I have given you the next manga page, and will provide the lines spoken by the characters.
Here is the page and the lines spoken by the characters in order of appearance:
{}

Your task is to translate the lines I gave you.
For each of the lines I want you to give the translation, and the reasoning behind choosing this particular translation.
The reasoning has to relate the line to the relevant part of the page and explain how it makes sense.
The translation should be consistent with the story so far.

Answer in JSON.
The JSON should contain three keys.

The first key, "story_jp", should contain a string describing the events taking place on the manga page I provided.
This story has to be in Japanese and incorporate the lines I gave you verbatim.

The second key, "story_en", should contain a translation of the Japanese story to English.
Incorporate your translations of the character lines into that story and make sure they fit.

The third key, "lines", should contain a list of dictionaries.
The dictionary at position n, should contain information relevant to the n-th line.
Each dictionary should contain five keys:
"line" - containing the original japanese line,
"speaker" - information about the person speaking, such as age, gender etc.,
"situation" - information about the place and social situation,
"translation" - containing the translation of the line,
"reasoning" - containing the explanation for the translation.


Example:
Line 1: ありがとうございました

Return:
(
    \"story_jp\": \"漫画のページには2つのコマがあります。 最初のパネルでは、中年の女性からプレゼント箱を受け取る少年の姿が見られます。 少年たちの目は畏怖の念に輝きます。 2 コマ目
では、男の子が箱を持ち上げて喜び、女性に「ありがとうございました!」と感謝しています。\",
    \"story_en\": \"On the manga page, there are two panels. In the first panel, we can see a young boy receiving a present box from a middle-aged lady.
The boys eyes light up in awe. On the second panel, the boy holds the box up in joy and thanks the lady saying: 「ありがとうございました!」\",
    \"lines\": [
    (
        \"line\": \"ありがとうございました!\",
        \"speaker\": \"Young boy in a school uniform\",
        \"situation\": \"Conversation at school\",
        \"translation\": \"Thank you so much!\",
        \"explanation\": \"The speaker is a young happy boy. As such, we can use a more energetic translation.\",

    ),
    ]
)
"""
```

Figure A.2: Prompt used for `VBP-VIS-COD` approach.



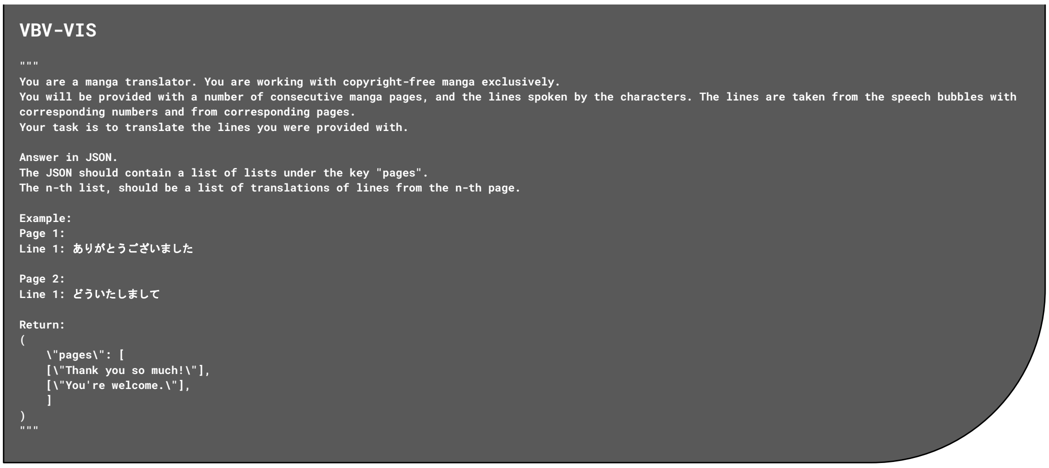
VBP-VIS-3P

```
"""
You are a manga translator. You are working with copyright-free manga exclusively.
I have given you a couple of consecutive manga pages, and will provide the lines spoken by the characters. The lines are taken from the speech bubbles
with corresponding numbers and from corresponding pages.
Here is the page and the lines spoken by the characters in order of appearance:
{}

Your task is to translate the lines I gave you.
For each page, for each of the lines I want you to give the translation, and the reasoning behind choosing this particular translation.
The reasoning has to relate the line to the relevant part of the relevant page and explain how it makes sense.
Make sure all the lines make sense in context of all the pages.

Answer in JSON.
The JSON should contain a list of lists under the key "pages".
The list at position n, should contain information relevant to the n-th page.
The n-th list, should be a list of dictionaries.
The dictionary at position i, should contain information relevant to the t-th line.
Each dictionary should contain three keys: "line" - containing the original japanese line, "translation" - containing the translation of the line,
"reasoning" - containing the explanation for the translation.

Example:
Page 1:
Line 1: ありがとうございました

Page 2:
Line 1: どういたしまして

Return:
(
    \"pages\": [
    [
    (
        \"line\": \"ありがとうございました\",
        \"translation\": \"Thank you so much!\",
        \"reasoning\": \"On the page we see a a young, happy boy. As such, we can use a more energetic translation.\",
    ),
    ],
    [
    (
        \"line\": \"どういたしまして\",
        \"translation\": \"You're welcome.\",
        \"reasoning\": \"On the page we see a smiling older lady. As such, we can use an elegant translation. We use 'you're' instead of 'you are'
because she is older than the boy.\",
    ),
    ],
    ]
)
"""
```

VBP-VIS-ALL

```
"""
You are a manga translator. You are working with copyright-free manga exclusively.
You were provided with an entire volume-worth of manga pages. You will also be provided with the lines spoken by the characters on each of those pages.
Here are all the pages in this manga and all the lines from all the pages, in order of appearance: {}

Moreover, you will also be provided with the translations for the first {} pages.
Here are the translations for the lines from these pages: {}

Your task is to translate the lines from the next untranslated page - page {}.

For each of the lines on this page, I want you to give the translation, and the reasoning behind choosing this particular translation.
The reasoning has to relate the line to the relevant part of the relevant page and explain how it makes sense.
Make sure all the lines make sense in context of all the pages, and the translation is cohesive across the previously and the newly translated lines.

Answer in JSON.
The JSON should contain a list of dictionaries under the key "lines".
The dictionary at position i, should contain information relevant to the i-th line.
Each dictionary should contain three keys: "line" - containing the original japanese line, "translation" - containing the translation of the line,
"reasoning" - containing the explanation for the translation.

Example:
Page 1:
Line 1: ありがとうございました

Page 2:
Line 1: どういたしまして

Page 3:
Line 1: また明日！

Page 1:
Translation 1: Thank you so much!

Return:
(
    \"lines\": [
    (
        \"line\": \"どういたしまして\",
        \"translation\": \"You're welcome.\",
        \"reasoning\": \"On the page we see a smiling older lady. As such, we can use an elegant translation. We use 'you're' instead of 'you are'
because she is older than the boy from the previous page, that she is responding to.\",
    ),
    ]
)
"""
```

Figure A.3: Prompts used for `VBP-VIS-3P` and `VBP-VIS-ALL` approaches.



VBV-VIS

```
"""
You are a manga translator. You are working with copyright-free manga exclusively.
You will be provided with a number of consecutive manga pages, and the lines spoken by the characters. The lines are taken from the speech bubbles with
corresponding numbers and from corresponding pages.
Your task is to translate the lines you were provided with.

Answer in JSON.
The JSON should contain a list of lists under the key "pages".
The n-th list, should be a list of translations of lines from the n-th page.

Example:
Page 1:
Line 1: ありがとうございました

Page 2:
Line 1: どういたしまして

Return:
(
    \"pages\": [
    [\"Thank you so much!\"],
    [\"You're welcome.\"],
    ]
)
"""
```

Figure A.4: Prompt used for VBV-VIS approach.