Context-Informed Machine Translation of Manga using Multimodal Large Language Models

Philip Lippmann♣* **Konrad Skublicki**♣ **Joshua Tanner**♢
Shonosuke Ishiwatari♢ **Jie Yang**♣
♣ Delft University of Technology ♢ Mantra Inc.

Abstract

Due to the significant time and effort required for handcrafting translations, most manga never leave the domestic Japanese market. Automatic manga translation is a promising potential solution. However, it is a budding and underdeveloped field and presents complexities even greater than those found in standard translation due to the need to effectively incorporate visual elements into the translation process to resolve ambiguities. In this work, we investigate to what extent multimodal large language models (LLMs) can provide effective manga translation, thereby assisting manga authors and publishers in reaching wider audiences. Specifically, we propose a methodology that leverages the vision component of multimodal LLMs to improve translation quality and evaluate the impact of translation unit size, context length, and propose a token efficient approach for manga translation. Moreover, we introduce a new evaluation dataset – the first parallel Japanese-Polish manga translation dataset – as part of a benchmark to be used in future research. Finally, we contribute an open-source software suite, enabling others to benchmark LLMs for manga translation. Our findings demonstrate that our proposed methods achieve state-of-the-art results for Japanese-English translation and set a new standard for Japanese-Polish.[1]

1 Introduction

A Japanese style of comics – referred to as *manga* – has been popular with audiences outside of Japan for decades. Handcrafting high quality translations, key to distributing manga world wide, is a difficult undertaking that takes significant time and effort. As such, most manga never leave the domestic Japanese market. Additionally, readers who do not speak a language into which manga is typically translated have limited or no access at all due to the high initial costs of translations.

The use of Neural Machine Translation (NMT) promises seamless translations from one language to another without involving a human translator (Sutskever et al., 2014; Vaswani et al., 2017). Still, successful applications of NMT to manga – or comics in general – remain limited, and automatic methods remain far from being able to reliably translate manga at a level comparable to humans (Hinami et al., 2021). This is in part due to the unique requirements of manga as a translation problem, which involves literary translation, handling split sentences across multiple speech bubbles, and especially resolving ambiguities using visual information. For example, in figure 1, achieving an accurate translation requires integrating both textual and visual context from the current and preceding scenes.

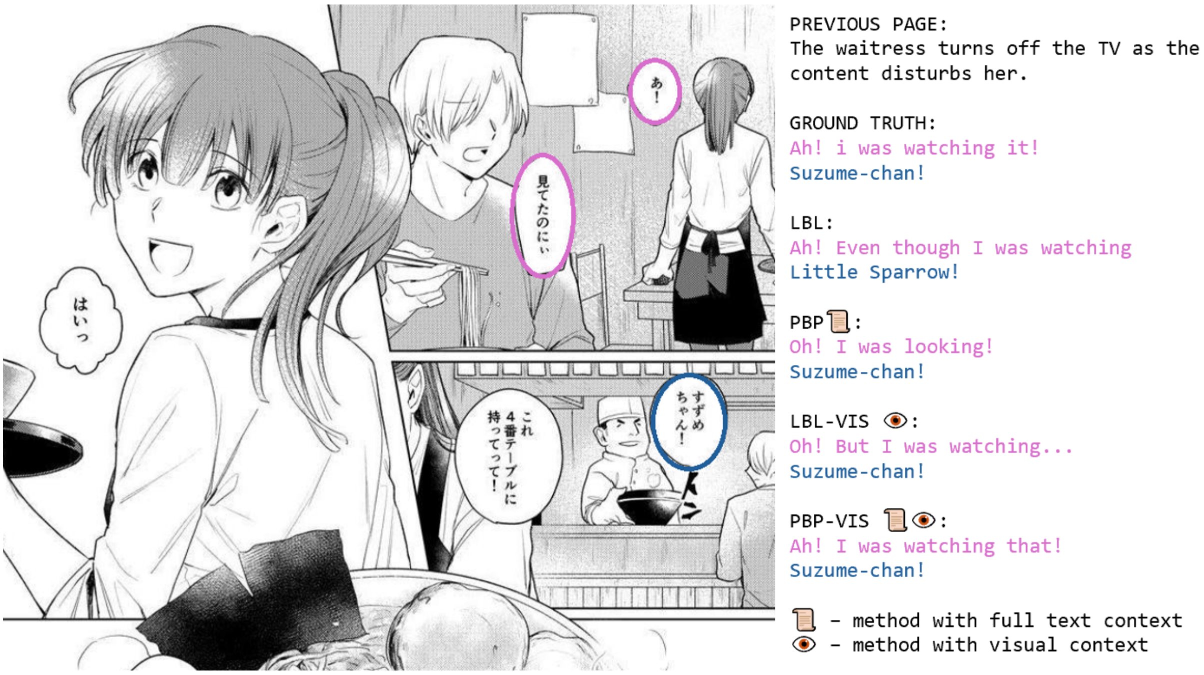
Research into manga-specific NMT methods is limited, focusing mainly on Japanese-English translation due to a lack of parallel corpora for other language pairs (Hinami et al., 2021; Kaino et al., 2024). Of these, only one method has attempted incorporating visual context into a model via a limited number of descriptive tags, yielding inconclusive results (Hinami et al., 2021). Previously proposed models were trained on a private JA-EN data set, which is not shareable due to copyright (Hinami et al., 2021; Kaino et al., 2024). Although there exist several general purpose manga data sets, such as Manga109 (Fujimoto et al., 2016; Matsui et al., 2017), so far only one manga translation data set has been published for research purposes: OpenMantra (Hinami et al., 2021). However, its limited size makes it effectively an evaluation data set only, making it challenging to train models.

Large language models (LLMs) have shown to be capable translators across languages (Lyu et al., 2023; Hendy et al., 2023). The release of multimodal LLMs – those that make use of visual infor-

[1] Data and code available at: `https://github.com/plippmann/multimodal-manga-translation`.

*All correspondence: `p.lippmann@tudelft.nl`.

Proceedings of the 31st International Conference on Computational Linguistics, pages 3444–3464
January 19–24, 2025. ©2025 Association for Computational Linguistics



PREVIOUS PAGE:
The waitress turns off the TV as the content disturbs her.

GROUND TRUTH:
Ah! i was watching it!
Suzume-chan!

LBL:
Ah! Even though I was watching
Little Sparrow!

PBP📜:
Oh! I was looking!
Suzume-chan!

LBL-VIS 👁:
Oh! But I was watching...
Suzume-chan!

PBP-VIS 📜👁:
Ah! I was watching that!
Suzume-chan!

📜 - method with full text context
👁 - method with visual context

Figure 1: Comparison of translation outputs for methods with different context types. The preceding scene visually shows a TV, giving context to the complaints in purple. The previous and current scenes are set in a restaurant, making it improbable that "Suzume" refers to a sparrow rather than being a name. ©Kira Ito

mation in addition to text – makes translation of media with visual nuance a possibility (Lyu et al., 2024). This potentially bypasses the need for large parallel manga corpora for each language pair as LLMs do not need to be finetuned by the user. Still, it is not clear how to best use these as effective manga translators.

In this paper, we present a translation methodology using a multimodal multilingual LLM. We evaluate a range of LLM-based translation approaches to empirically assess the impact of visual context, translation unit size, and context length. We do this using an existing JA-EN manga data set and a new Japanese-Polish data set created for this purpose. The JA-PL translation direction is chosen for its unique challenges, particularly due to the significant differences in syntactic structures and semantic nuances between Japanese, English, and Polish, and to address a low-resource language that nonetheless has a market for manga (Świeczkowska, 2017). Finally, we contribute an open-source manga translation evaluation suite that allows users to choose the granularity of available context, provides automatic evaluation metrics, and enables testing of different LLMs.

In summary, our contributions are as follows:

- An LLM-based multimodal manga translation methodology that achieves state of the art results on JA-EN translations and can serve as a baseline for low-resource languages.
- An annotated set of 400 professionally translated manga pages (3705 sentences) that make up the first ever parallel JA-PL manga translation benchmark data set, as well as the largest manga translation data set to date.
- The first publicly available automatic manga translation evaluation software suite.

2 Related Work

2.1 Automatic Methods for Manga

Up to this point, the development of automatic manga translation methods that incorporate multimodal context has been limited. Hinami et al. (2021) first proposed an NMT system for manga that makes use of contextual information obtained from images to inform the translation. Their method is restricted to a single frame of context and the visual information obtained from the images is limited to 512 predefined labels. Further work has explored the use of an additional frame or manga metadata to improve translation quality (Kaino et al., 2024), however, without visual context. Instead, we propose taking additional textual content of up to the entire manga volume into account to improve translations, as well as the full manga image without predefined labels. Outside of translations, Chen et al. (2019) propose a sentiment analysis method on manga text and Guo et al. (2023a) propose an approach that makes use of both visual and textual modalities to complete empty speech bubbles in existing manga. There has been sparse early-stage research into automatic methods for similar media, such as graphical novels (Harshavardhan et al., 2024) and American comics (Hapsani et al., 2017).

2.2 Large Language Model Translations

Translation using LLMs is appealing due to their ability to generate high-quality translations for various language pairs without the need for training on extensive parallel corpora or fine-tuning (He et al., 2023). LLMs have previously been shown to be capable translators (Wang et al., 2023), as well as evaluators of translation quality (Kocmi and Federmann, 2023). Further, paragraph translations performed by LLMs have been shown to be effective when using basic English prompts at the sentence level (Zhang et al., 2023a). We propose multiple translation approaches and evaluate the quality of our LLM manga translations compared to finetuned transformer models and explore a low-resource language pair, JA-PL, as well as contribute a data set for evaluation.

2.3 Multimodal Machine Translations

Translating text embedded in images has been extensively explored in research (Zhu et al., 2023; Lan et al., 2023). Multimodal machine translation (MMT) has so far mainly been applied to translating image captions, outperforming the text-only baseline by leveraging additional visual information (Gwinnup and Duh, 2023). MMT typically uses a single image with its corresponding text description as input (Elliott et al., 2016). We investigate to what extend an increased visual context length is effective. A further challenge comes from the discrepancy between the natural images and their description used to train vision encoders used for MMT and manga images, as manga has a unique hand-drawn art style with relevant text drawn into the image (Guo et al., 2023b). Additionally, little attention has been paid to low-resource languages, with the vast majority of MMT research focused on the most popular language translation pairs (Guo et al., 2022; Huang et al., 2023b). More recent LLMs have additional multimodal capabilities (Huang et al., 2023a; Yin et al., 2023), enabling them to perform MMT, though this has not been explored for the manga use case.

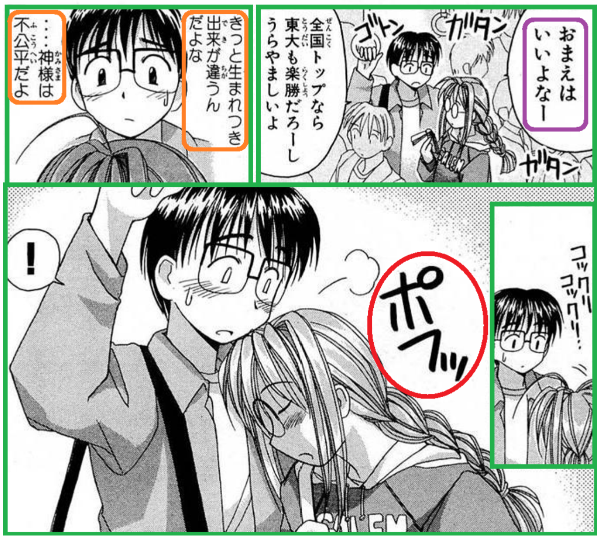

おまえはいいよなー
ガタン
ガタン
ゴトン
全国トップなら東大も楽勝だろーしうらやましいよ
きっと生まれつき出来が違うんだよな
……神様は不公平だよ
コックリ
コックリ!!
ポフッ
!

Figure 2: A manga page: panel borders (**green**), example lines in speech bubbles (**purple**), free flowing text (**orange**) and sound effects (**red**). Courtesy of Akamatsu Ken, ©Kodansha

3 Methodology

In this section, we first outline manga terminology, then present the problem, and finally introduce our LLM-based translation approaches that take advantage of multimodality and a longer context.

3.1 Terminology & Problem Formulation

Page-to-page manga translation involves three steps: (1) *page processing* to identify elements on the page, detect text, and estimate reading order; (2) *translating* the text into the target language; and (3) *typesetting* the translated text onto the page in stylized font. The focus of this paper is on (2), but we will discuss (1) and (3) in appendix A.

A manga page consists of multiple story panels, referred to simply as *panels*, as shown in Figure 2. Panels often contain text, which can be enclosed in a *speech bubble* for text spoken or thought by characters, or free-flowing for background noise or sound effects. The term *line* will always refer to the content of one speech bubble, narrative box, or cluster of free-flowing text.

For multimodal manga translation, we make use of the *image* of the drawings on a single manga page, such as figure 2, which contains lines of text. We make the assumption that the text contained on the page has already been recognized and is available. Our goal is to obtain the correctly translated text for each line.

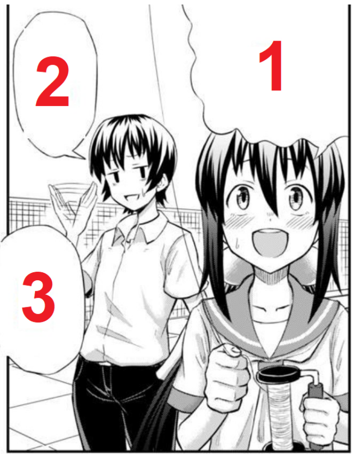


Figure 3: Fragment of a page annotated for the PBP-VIS-NUM method ©Mitsuki Kuchitaka.

3.2 Translation Approaches

We use a variety of translation approaches – summarized in table 1 – to assess the impact of multimodality, translation unit size, and context length and find the most performant configuration. To establish a baseline, our first approach is a simple line-by-line approach (LBL). This means that the model receives one single line to translate at a time, without any additional context about the manga it is translating. Previous research has shown that LLMs perform better on translation tasks when given the entirety of a text compared to snippets, as they are able to incorporate the broader context more effectively (Karpinska and Iyyer, 2023). As such, the second approach we evaluate is page-by-page (PBP), where the model is given all lines from a given page in the correct reading order and outputs all the corresponding translations.

3.2.1 Multimodal Translation

Ideally, we would provide the LLM with just the image and it would recognize the text, obtain the visual context, and perform the translation. However, current models are not capable of this. Instead, we investigate approaches where the model is given the lines on a page to translate, along with the corresponding image as additional visual context, enabling multimodal translation.

The first approach we investigate is the multimodal equivalent to LBL, referred to as LBL-VIS, where the model receives lines and the corresponding page image as visual context. The second approach utilizing visual context is PBP-VIS, which involves giving the model the entire text from one manga page and the page itself as an image.

Approach	Translation Unit	Textual Context	Visual Context
LBL	line	line	×
PBP	page	page	×
LBL-VIS	line	line	page
PBP-VIS	page	page	page
PBP-VIS-NUM	page	page	num. page
VBP-VIS-COD	page	page + sum.	page
VBP-VIS-3P	page	3 pages	3 pages
VBP-VIS-ALL	page	vol. + trans.	volume
VBV-VIS	volume	volume	volume

Table 1: Overview of the proposed approaches. Abbreviations: "num." is numbered, "sum." is summary, "vol." is volume, and "trans." is translation so far.

The final approach aims to directly address the issues that multimodal LLMs have with reading non-Latin scripts. The setup is the same as PBP-VIS, but the image of the manga page is modified to avoid the LLM performing any optical character recognition (OCR). The contents of the speech bubbles in the image 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 figure 3). We call this approach PBP-VIS-NUM and it enables the model to locate the speech bubble more easily and relate its content to the exact panel in which it was placed, without performing any OCR on the text itself.

3.2.2 Long-Context Translation

Intuitively we want to make use of context lengths exceeding single lines or pages to adequately capture evolving story lines and character development and accurately translate entire stories in an internally consistent way. The remaining approaches we present are designed to address this.

The first of these multi-page approaches provides the model with the previous and next page as additional context to give more information to the LLM. We refer to this as VBP-VIS-3P, as it translates the volume sequentially one page at a time (VBP), using visual context (VIS), and using three pages' worth of context (3P). Going a step further, we explore VBP-VIS-ALL, where the model is provided with the images and lines from an entire manga volume, as well as all the translations so far, and queried to translate the next untranslated page. This process is repeated sequentially for every page in the volume.

As the input and output length increases, the limited context windows of LLMs are quickly exhausted and performance is diminished (Liu et al.,

Manga Title	Genre	# Pages	# Lines
Balloon Dream	Romance	38	314
Boureisougi	Mystery	36	274
Rasetugari	Fantasy	54	359
Tencho Isoro	SoL	40	311
Tojime no Siora	Battle	46	334

Table 2: Overview of the OpenMantra data set (Hinami et al., 2021). SoL is slice-of-life.

2023). To overcome this, we introduce the scalable VBP-VIS-COD approach, where we extend chain of density summarization (COD) (Adams et al., 2023) to keep a rolling, fixed-length summary of the story's developments as our context. Besides the image and its corresponding text, the model is given a summary of the story thus far in the target language as additional context. For a detailed overview of this approach, see appendix B. The last evaluated approach, VBV-VIS translates an entire manga volume in a single call. Similar to VBP-VIS-ALL, we provide the LLM with the texts and images from an entire manga volume, but then instruct it to respond with the translations for the entire volume in a single query.

4 Experiments

4.1 Data

Machine translation of manga is a niche field with little publicly available data and no established research benchmarks (Hinami et al., 2021). There exists a plethora of manga data that could be used for training and evaluating machine translation systems (Świeczkowska, 2017; Sachdeva and Zisserman, 2024). However, from a research perspective, the main issue with manga is that, due to its commercial nature, most manga is protected by Japanese and local copyright laws (Schroff, 2019). Previous manga-related works have addressed this problem in different ways. Some researchers resort to using private data sets (Rigaud et al., 2021; Hinami et al., 2021; Kaino et al., 2024), while others use the very few publicly available copyright-free manga, accepting the trade-off of unlabeled data (Sharif et al., 2021).

To date, there has been only one manga translation data set made public for research purposes – the OpenMantra data set by Hinami et al. (2021). It consists of five independent Japanese-language manga volumes, totaling 214 pages (1593 speech bubbles). Details of this data set are shown in table 2. Each volume in this data set 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 data set to evaluate JA-EN translations, splitting it into two parts: validation set (*Balloon Dream* and *Tojime no Siora*) and test set (*Boureisougi*, *Rasetugari*, and *Tencho Isoro*).

4.2 New Japanese-Polish Manga Data Set

In addition to JA-EN, we explore JA-PL translation; as English and Polish belong to different language families, diverge significantly in terms of morphology, and have different grammatical structures. We provide professional JA-PL translations of the slice-of-life manga *Love Hina* to create a data set for research purposes. We make volumes 1 and 14 available and our annotation process closely follows the existing annotations of the Japanese text. The newly contributed data set contains 400 pages and 3705 individual lines (*i.e.* speech bubbles, sound effects, etc.) split across the two volumes and is distributed as a set of images, corresponding to one image per page, and the corresponding metadata containing original and translated text, as well as their coordinates on the page. This exceeds the previously largest manga translation data set, OpenMantra (Hinami et al., 2021), in size. We propose a 50:50 validation:test split for this data set, using the first volume (200 pages and 1810 lines) as the test set and the last volume (200 pages and 1895 lines) as the validation set. This decision is motivated primarily by the fact that the first volume establishes the story, providing a fairer benchmark for the long-context methods, as opposed to the last volume, which depends on unavailable context, being the 14th installment in the series.

Our annotation process closely follows the existing annotations of the Japanese text. The original lines were matched with the corresponding translated lines primarily based on location, and if impossible, based on content. However, in edge cases the Polish edition left very small text untranslated as a stylistic choice. The reading order was first estimated using the tool provided by Sachdeva and Zisserman (2024) and then corrected by hand based on the actual speech bubbles. During the annotation process, we noticed several characteristics of this title and the unique challenges it presents for translation. Some characters in *Love Hina* speak the Kansai dialect of Japanese. According to the literature, there is no consensus on how to translate this dialect into Polish, with different translators choosing different Polish dialects (Jaśkiewicz, 2020). Another challenge is that one of the secondary characters speaks in a manner resembling samurai speech – a common trope in manga (Duc-Harada, 2019). Again, there is no consensus on how to convey this in Polish. As such , users of the data set should be aware that some "incorrect" translations may be just as valid in these cases.

4.3 Baselines

We employ four baseline methods for JA-EN translations. The first two baselines, `Scene-NMT` and `Scene-NMT-VIS`, come from the original automatic manga translation work by Hinami et al. (2021). The first method uses a transformer-based model to translate the contents of entire panels at once without multimodal context, while the second method includes visual features as well. The third baseline method we use – and current state-of-the-art for automatic manga translation – is `Scene-aware-NMT` (Kaino et al., 2024), which translates manga panel by panel as well, using a transformer-based model but uses the text from the previous panel as additional context. The translation outputs for all these previously listed methods were kindly provided to us by the authors of the respective works. This allowed us to use our own data splits and ensure that all methods were evaluated equally and comparably.

The last baseline method we use is Google Translate (`GT`) due to its support for a wide range of languages and availability. `GT` is our only baseline for JA-PL translations. All `GT` translations were carried out in April and May 2024, using the Google Translate API with the corresponding Python library.[2]

4.4 Automatic Evaluation

For evaluation, we use a range of automated metrics applied at the sentence level. We use a lexical n-gram matching heuristic metric in ChrF (Popović, 2015). Although the reliability of this type of metric has been questioned over the years (Thai et al., 2022), they remain among the most widely used in machine translation (Mathur et al., 2020; Kocmi et al., 2024). ChrF provides scores on a scale from 0 to 100, where higher scores indicate higher quality translations.

The first non-lexical machine translation evaluation metric we use is BERTScore (Zhang et al., 2019), considered a good representative of the embedding-based metrics category (Saadany and Orasan, 2021). Although not perfect, it has been shown to detect important content words and is well suited to score candidates from different languages (Hanna and Bojar, 2021). Next, we report scores with a learned metric, BLEURT (Sellam et al., 2020), specifically the top-performing BLEURT-20 model (Pu et al., 2021). The last metric we report is the learned metric xCOMET (Guerreiro et al., 2023), specifically xCOMET-XXL. xCOMET is an open-source learned metric that performs error span detection in addition to standard sentence-level evaluation. It is currently considered the best-performing publicly available metric (Freitag et al., 2023). Among all the metrics we employ, it is the only one that calculates its score based not only on the references and hypotheses but also on the source text. BERTScore, BLEURT, and xCOMET return a score on a scale of 0 to 1, with results closer to 1 being preferable.

4.5 Human Evaluation

In addition to our extensive automatic evaluation, we perform a human evaluation with a professional JA-EN manga translator using the Multidimensional Quality Metrics (MQM) translation evaluation framework (Burchardt, 2013; ISO 5060:2024). We use MQM with a manga-specific list of issue types that cover different types of errors, such as accuracy, fluency, and style. A complete overview of our MQM process is shown in appendix C. Each error type is assigned a severity level, ranging from minor to critical, depending on the impact of the issue on overall quality. MQM provides a scoring system that allows for the calculation of overall quality scores based on the number of identified issues and their severity levels. These scores have an upper bound of 1 and no lower bound, with a higher score being preferable. We choose the *Tencho Isoro* manga for our MQM evaluation. We compare the official commercial translation of the manga, the `GT` baseline, and our best performing approach (`PBP-VIS`) to evaluate how a professional human translator would judge each.

4.6 Prompting

We follow the approach of previous works (Hendy et al., 2023; Karpinska and Iyyer, 2023; Lyu et al., 2024) and investigate the out-of-the-box translation performance of GPT-4 Turbo (OpenAI et al., 2024). The specific version we use is

[2] https://pypi.org/project/googletrans/

Method	🇯🇵 **JA-EN** 🇺🇸				🇯🇵 **JA-PL** 🇵🇱			
	ChrF	**BRTS**	**BLRT**	**xCMT**	**ChrF**	**BRTS**	**BLRT**	**xCMT**
`GT`	34.2	0.895	0.525	0.729	22.3	0.826	0.446	0.457
`Scene-NMT`	34.2	0.897	0.512	0.651	-	-	-	-
`Scene-NMT-VIS`	34.5	0.895	0.507	0.664	-	-	-	-
`Scene-aware-NMT`	36.1	**0.903**	0.534	0.670	-	-	-	-
`LBL`	32.7	0.883	0.523	0.716	24.2	0.844	0.495	0.531
`PBP`	36.0	0.898	0.565	0.758	25.6	**0.852**	0.538	0.565
`LBL-VIS`	35.6	0.900	0.551	0.746	24.9	0.845	0.515	0.543
`PBP-VIS`	36.6	**0.903**	0.581	**0.776**	25.6	**0.852**	**0.539**	**0.567**
`PBP-VIS-NUM`	**36.8**	0.900	**0.582**	**0.776**	**25.7**	0.851	0.532	0.566
`VBP-VIS-COD`	35.9	0.900	0.566	0.769	25.1	0.846	0.523	0.550
`VBP-VIS-3P`	35.9	0.897	0.565	0.754	25.6	0.843	0.530	0.559
`VBP-VIS-ALL`	35.7	0.893	0.556	0.760	24.9	0.840	0.521	0.561
`VBV-VIS`	34.9	0.884	0.539	0.733	24.5	0.833	0.510	0.534

Table 3: Performance metrics for all approaches for JA-EN and JA-PL translation. Best scores for each translation direction are in **bold**. BRTS refers to BERTScore, BLRT to BLEURT, and xCMT to xCOMET.

`gpt-4-turbo-2024-04-09` at default hyperparameters with a temperature $T = 0.5$, accessed through the OpenAI Python library.[3] For all multimodal translations, we append the relevant image(s) of the page(s) as a *jpeg* file to the LLM query via its respective API. We run each configuration once due to the high costs involved in sending entire manga volumes to commercial multimodal LLMs. The complete prompts we use for every translation are shown in appendix D. Each approach described in section 3 is evaluated one-shot, i.e., with one given example in the prompt. We did not find a measurable difference between one-shot and five-shot prompting when evaluating on the validation data. The model is always prompted in English – regardless of the target language – as this has yields the best results for LLM translations (Zhang et al., 2023b). Based on experiments on the validation data, we ask the model to explain how the image influences the translation, ensuring that the visual context is taken into account.

4.7 Manga Translation Evaluation Suite

We release our evaluation suite to advance research in automatic manga translation. It enables benchmarking of various LLMs by adjusting textual context, visual context, and translation unit size. The suite integrates all methods from section 3 for comprehensive evaluation and facilitates automatic assessment using the four previously outlined metrics. With plug-and-play functionality, researchers can easily utilize existing data sets, including Open-Mantra and ours, while introducing new prompts and exploring alternative LLMs.

5 Results & Discussion

5.1 JA-EN Translation

We present our findings in table 3. Among the methods proposed in previous studies, `Scene-aware-NMT` demonstrates competitive performance, especially on BERTScore, surpassing other previous manga-focused translation methods, consistent with their reported findings. However, our proposed methods show improvements across multiple metrics. Our basic approach, `LBL`, performs slightly worse than `GT` in most aspects. The `PBP` method shows substantial improvement over `LBL`, outperforming all baselines on BLEURT (0.565) and xCOMET (0.758), confirming the potential of LLMs as manga translators, even without visual context.

Visual Context The addition of visual context significantly improves scores across all metrics for both `LBL` and `PBP` methods. `PBP-VIS` and `PBP-VIS-NUM` achieve the best scores across most metrics, with `PBP-VIS-NUM` slightly outperforming on ChrF (36.8) and BLEURT (0.582). These results confirm that additional visual context significantly

[3] https://github.com/openai/openai-python

improves LLM translation quality, representing a novel approach in automatic manga translation. Additionally, we perform an ablation study to clarify the role of key visual features, the results of which are discussed in appendix E.

The results of our human evaluation are presented in table 4. PBP-VIS, clearly outperforms the GT baseline in overall score. However, according to the MQM evaluation conducted by a single professional translator, PBP-VIS is more prone to errors than the official human manga translation. While our method has fewer "minor" and "major" errors compared to the official translation, it exhibits a significantly higher number of "critical" errors. These findings indicate that although our method establishes the current state of the art for automatic manga translation, human translation remains superior in quality. Although these metrics provide context for assessing our method's efficacy, translation quality is inherently subjective and challenging to measure. While our best translation scores lower than the official translation, we find it enjoyable to read and coherent – a standard the GT translation does not meet.

Context Length Interestingly, providing context beyond the page level does not enhance translation quality. VBP-VIS-COD, using only a short summary of previous events, performs better than other long-context methods across most metrics. Conversely, VBV-VIS, which translates the entire volume in one query, shows the lowest performance among our visual context methods. These findings suggest an inverse relationship between translation quality and input length beyond a single page for multimodal LLM translation. This counter intuitive result highlights the importance of optimizing input size for LLM-based translations.

5.2 JA-PL Translation

For the JA-PL data, we do not report the results of methods proposed by other authors, as these are not trained on Polish data and therefore perform poorly. For JA-PL translations, we observe that across methods, scores are generally lower compared to JA-EN translations. However, all our approaches significantly outperform the GT baseline. Further, we note that our top performing methods, PBP-VIS and PBP-VIS-NUM, perform similarly on JA-PL to JA-EN.

Visual Context and Context Length Again visual context improves performance, though to a much lesser extend than for JA-EN translation. For PBP, the impact of visual context is minimal, with PBP-VIS and PBP-VIS-NUM performing similarly to PBP. Long-context approaches show mixed results, again performing worse than PBP-VIS and PBP-VIS-NUM.

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

Table 4: Human evaluation MQM results for JA-EN. Errors are number per category (lower preferable).

5.3 Implications and Broader Impact

PBP-VIS and PBP-VIS-NUM consistently achieve the best results for both JA-EN and JA-PL translations. The effectiveness of our methods across translations suggests broad applicability to different language pairs. The cross-lingual success of our methods indicates that the benefits of incorporating visual context in manga translation are language-independent. Moreover, our PBP-VIS and PBP-VIS-NUM methods achieve the highest scores across all metrics, setting the state of the art for automatic manga translation.

Notably, we observe that translation quality does not necessarily improve with longer context, challenging common assumptions in machine translation. This finding aligns with previous research, which indicates that the quality of output from LLMs tends to diminish as the length of the input increases (Levy et al., 2024). The is contrary to the results we observe when additional visual context is taken into account. To optimize performance when using LLMs for multimodal translations, it is advisable to prioritize smaller input sizes of a single page. Translation quality tends to deteriorate more significantly as the LLM processes longer text, even if it contains more information relevant to the story.

6 Conclusion

Our investigation of multimodal LLMs for automatic manga translation reveals significant advancements in this emerging field. We evaluate various LLM-based translation approaches, considering text-only, image-informed, and volume-level contexts. Leveraging the vision component of multimodal LLMs, we enhance translation quality by incorporating visual elements to resolve ambiguities. However, we find that additional textual context does not consistently improve performance. Our methodology achieves state-of-the-art results for JA-EN translations and sets a new standard for JA-PL translations. We also introduce the first parallel JA-PL manga translation data set and an open-source benchmarking suite for LLMs.

7 Limitations

The first limitation of this study is the amount of data used for testing. While we make meaningful contributions to addressing this issue, there is still a severe lack of evaluation data, making it difficult to determine how consistent our findings would be across different authors and genres. Additionally, we only investigate one language other than English, constrained by our ability to manually inspect outputs and analyze model mistakes in other languages.

Related to this is the fact that some manga series span multiple volumes. Translations of later volumes in a series would undoubtedly benefit from including earlier volumes in the available context. Due to the lack of suitable data, we limit ourselves to translations of single volumes, leaving multi-volume narratives to future work.

Finally, there are obvious limitations when using a commercial, closed-source LLM as we do in this paper, such as potential data leakage issues and the unlikely scenario that some of the manga used might have been part of the training data. Still, the availability and quality of open-source multimodal multilingual LLMs is very limited at this time, and as such we leave a study using alternatives to future work.

Acknowledgments

This work was supported by an Oracle for Research Grant Award, as well as SURF Grant EINF-8535.

References

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

Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. 2020. Building a manga dataset “manga109” with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18.

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

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Jiali Chen, Ryo Iwasaki, Naoki Mori, Makoto Okada, and Miki Ueno. 2019. 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.

Julián Del Gobbo and Rosana Matuk Herrera. 2020. 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.

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

Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. 2023. 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.

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

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

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongcheng Guo, Boyang Wang, Jiaqi Bai, Jiaheng Liu, Jian Yang, and Zhoujun Li. 2023a. M2C: Towards automatic multimodal manga complement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9876–9882, Singapore. Association for Computational Linguistics.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023b. Bridging the gap between synthetic and authentic images for multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2863–2874, Singapore. Association for Computational Linguistics.

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

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

Anggi Gustiningsih Hapsani, Fitri Utaminingrum, and Herman Tolle. 2017. Optical character recognition on english comic digital data for automated language translation. *Int. J. Advance Soft Compu. Appl*, 9(3).

G Harshavardhan, Sandeep Singh Kang, Geet Kiran Kaur, and Sanjay Singla. 2024. The future of graphic novel translation: Fully automated systems. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, volume 1, pages 1–8. IEEE.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *Preprint*, arXiv:2305.04118.

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

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. *Preprint*, arXiv:2012.14271.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Bjorck, Vishrav Chaudhary, Subhojit Som, XIA SONG, and Furu Wei. 2023a. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 72096–72109. Curran Associates, Inc.

Xin Huang, Jiajun Zhang, and Chengqing Zong. 2023b. Contrastive adversarial training for multi-modal machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

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

Hanna Jaśkiewicz. 2020. Reprezentacja dialektu bawarskiego i dialektu kansai w literaturze współczesnej w kontekście ideologii językowych w niemczech i japonii. In *Forum Filologiczne ATENEUM*.

Hiroto Kaino, Soichiro Sugihara, Tomoyuki Kajiwara, Takashi Ninomiya, Joshua B. Tanner, and Shonosuke Ishiwatari. 2024. Utilizing longer context than speech bubbles in automated manga translation. pages 17337–17342.

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

U-Ram Ko and Hwan-Gue Cho. 2020. Sickzil-machine: A deep learning based script text isolation system for comics translation. In *Document Analysis Systems*, pages 413–425, Cham. Springer International Publishing.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. *Preprint*, arXiv:2305.17415.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Preprint*, arXiv:2402.14848.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, Siyou Liu, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. *Preprint*, arXiv:2305.01181.

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

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

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. 2019. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Andrei Novikov. 2019. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230.

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 Radford, 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. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

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

Christophe Rigaud, Nhu-Van Nguyen, and Jean-Christophe Burie. 2021. 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.

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

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

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

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

Mhd Saeed Sharif, Bilyaminu Auwal Romo, Harry Maltby, and Ali Al-Bayatti. 2021. 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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

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

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13433–13447, Toronto, Canada. Association for Computational Linguistics.

A Page Processing and Typesetting

Page-to-page manga translation involves three distinct steps: (1) *page processing* to identify elements on the page, detect text, and estimate reading order; (2) *translating* the text into the target language; and (3) *typesetting* by removing the source text from the page and inserting the translated text in a stylized font. We will discuss (1) and (3) in this section.

Page Processing The first step in manga translation is identifying the elements on the page. Here, we will present an example of a manga page processing pipeline composed of methods proposed by previous research and publicly available manga tools. For text detection, we employ the unconstrained method proposed by Del Gobbo and Matuk Herrera (2020) to account for text that is not contained within speech bubbles – see the top right of figure 4. However, before applying Optical Character Recognition (OCR) to the detected text fields, we need to group it into clusters belonging to the same utterance. To accomplish this, we apply a method inspired by Rigaud et al. (2021) – we utilize the OPTICS algorithm (Ankerst et al., 1999), specifically the Python pyclustering library implementation (Novikov, 2019), to cluster the text – see the bottom left of figure 4. We then compute the bounding boxes of these obtained text clusters and discard those that are too small to contain text – see the bottom right of figure 4. Finally, we apply Manga OCR[4] for text recognition – see figure 5.

For panel detection and estimating the reading order, we utilize the Magi system (Sachdeva and Zisserman, 2024). In theory, Magi is capable of creating a transcript of a manga page independently, but it was trained on translations of manga and is

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

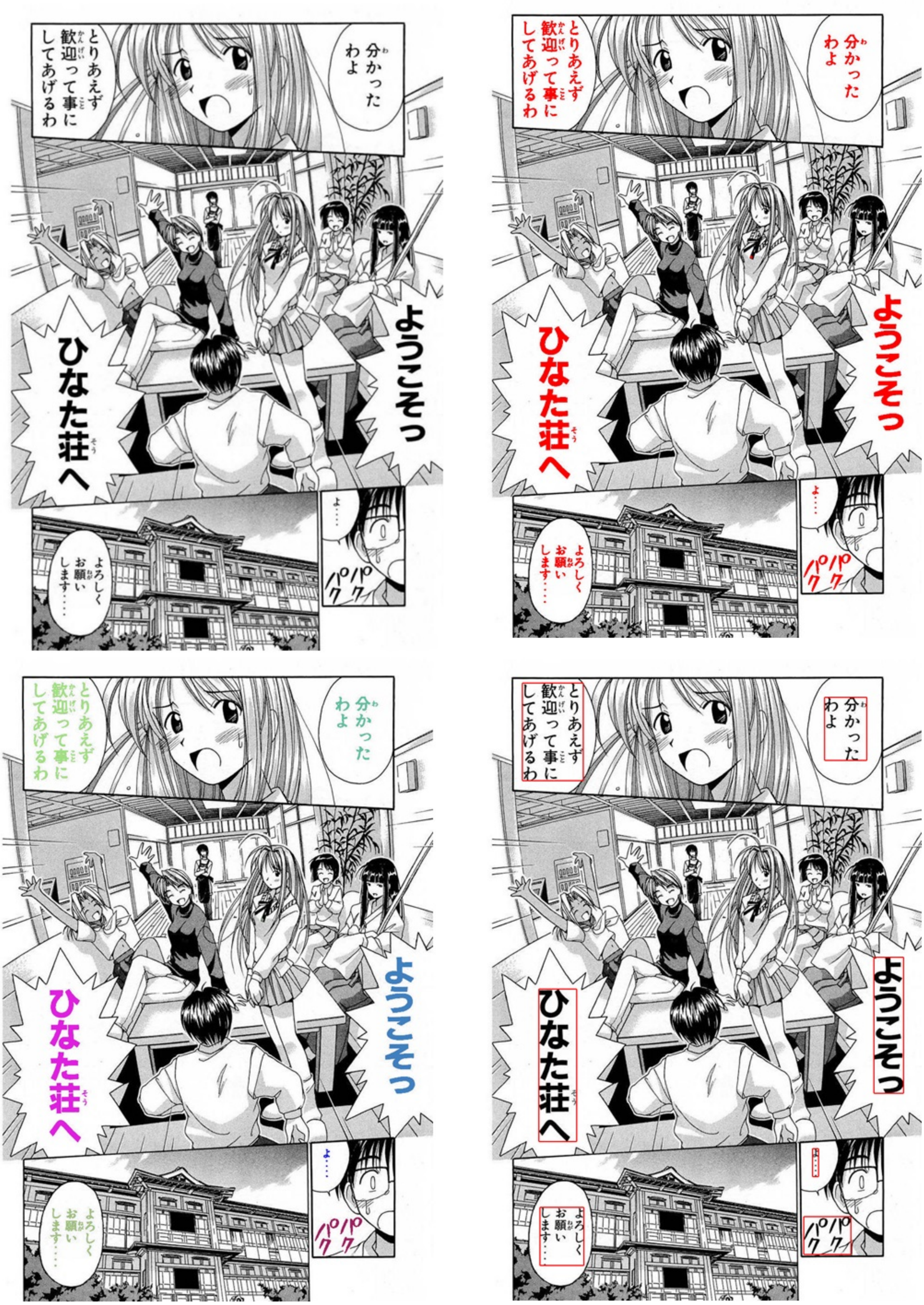
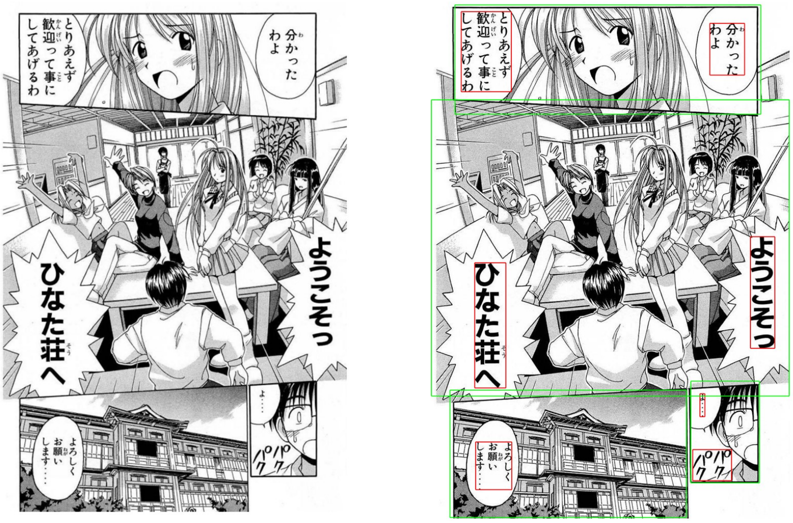


Figure 4: Stages of the text detection pipeline. First, pixels belonging to letters are identified. Then, the pixels are clustered into utterances. Lastly, bounding boxes are computed. Courtesy of Akamatsu Ken, ©Kodansha, from the Manga109-s dataset (Fujimoto et al., 2016; Matsui et al., 2017; Aizawa et al., 2020)



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 5: 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 (Fujimoto et al., 2016; Matsui et al., 2017; Aizawa et al., 2020)

not well-suited for Japanese text detection. As such, we only utilize some of its functionalities. A visualization of the page processing pipeline can be seen in Figure 5. First, we use the process described previously to detect text boxes. Then, we employ Magi to detect panels and estimate the reading order based on the relative locations of text boxes and panels. Lastly, we utilize Manga OCR for text extraction.

Typesetting The final step of a page-to-page manga translation – inserting the translated text back into the image – involves two steps: cleaning and lettering. We did not perform it as part of our study but, for the sake of giving potential future works a comprehensive guide to follow, we will still outline the procedure here. Cleaning refers to the removal step of the text, in which the original Japanese text is removed from the image used in the translation process. One could utilize an image inpainting model for this task (Nazeri et al., 2019); the regions containing text lines are replaced by the inpainting model, which effectively removes the text even when it is overlaid on textured images or drawings. Alternatively, a method that performs this step specifically for manga has been proposed (Ko and Cho, 2020).

Lettering is the final step, where the translated text is rendered onto the cleaned image, with an optimized font size and placement that fits the manga aesthetic. The location of the rendered text is chosen to maximize the font size while ensuring that all text remains within the designated text region. This step ensures that the translated text is legible and properly integrated into the image. To the best of our knowledge, there exists no prior work that proposes to do this automatically at this time. Though there are plenty of manual tools (both free and commercial) that make it possible to come up with a semi-automated approach if the coordinates of the original text are known.

B Chain of Density Summarization

In the context of text summarization, Chain of Density (COD) (Adams et al., 2023) prompting can be used to generate high-quality summaries by breaking down the summarization task into a series of smaller, more manageable sub-tasks. COD prompting can be applied to summarization as follows:
Initial Prompt: The process starts with a prompt that instructs the LLM to read and understand the source text that needs to be summarized.
First Generation: The LLM generates a brief summary or highlights the key points from the source text based on the initial prompt.
Iterative Prompting: The generated summary from the previous step is used as the prompt for the next step. The LLM is then prompted to expand or refine the summary by adding more details, rephrasing certain parts, or reorganizing the information. This step can be repeated multiple times, with each subsequent prompt building upon the previous summary.
Final Summary: After several iterations, the final summary should be a coherent, concise, and

```
Existing Summary from the previous Translation: {self.prev_context}

The most recent Observation from the {self.lang} translation was: {self.observation}

You will generate new increasingly concise, entity-dense summaries based on the above Existing
     Summary and most recent Observation.

Keep the summaries in {self.lang}.

You will create 3 summaries. You will create each of them by following the following two steps:
    -Step 1. If possible, identify 1-3 Informative Entities (";" delimited) from the most recent
         Observation which are missing from the Existing Summary.
    -Step 2. Write a new, denser summary of identical length which covers every entity, action, and
         detail from the previous Existing Summary plus the Informative Entities from the Observation.


An Informative Entity is:
    -Relevant: to the translation's unfolding narrative.
    -Specific: descriptive yet concise (10 words or fewer).
    -Novel: not in the previous summary.
    -Faithful: an accurate, detailed reflection of the translation.

Guidelines:
    -The first of the three summaries must be long (but less than ~{lmax} words) yet highly non-
         specific, containing little information beyond the entities marked as missing. Use verbose
         language and fillers (e.g., "In this part of the translation, the main character encounters
         ...") to reach ~{lmax} words.
    -Make every word count: rewrite the previous summary to improve flow and make space for
         additional Informative Entities.
    -Make every word count: rewrite the previous summary to improve flow and make space for
         additional Informative Entities.
    -Make space with fusion, compression, and removal of uninformative phrases like "the scenario
         presents".
    -The summaries should become highly dense and concise yet self-contained, e.g., easily understood
          without referencing the fact that a translation is being performed, and contain all
         information of the narrative thus far.
    -Informative Entities can appear anywhere in the new summary.
    -Only drop the least relevant Informative Entities from the previous summary if the summary
         length exceeds ~{lmax} words. Otherwise carry all previous Informative Entities to the new
         summary.

Answer in JSON. The JSON should be a list (length 3) of dictionaries under the key "summaries". Each
     dictionary should contain keys "Informative_Entities" (storing the Informative Entities included
      in the corresponding summary) and "Denser_Summary" (containing the summary).
```

Figure 6: Prompt used for Chain of Density summarization.

informative representation of the source text.

By breaking down the summarization process into smaller steps, COD helps the LLM maintain focus and context throughout the summary generation process. This can lead to more coherent and accurate summaries, as the model can incrementally refine and improve the summary at each step. It's important to note that the effectiveness of COD prompting for summarization may depend on the quality of the initial prompt, the complexity of the source text, and the LLM's capabilities.

The `VBP-VIS-COD` approach we propose does not try to make use of the large context window size of GPT-4 Turbo and instead uses COD to maintain a rolling summary of the developments in the story so far. In addition to 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 translation, 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 the LLM to update the previous summary with the new developments, to achieve an even denser and updated summary through the process detailed above, with each summary within the same call being a more concise version of the previous one. 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`).

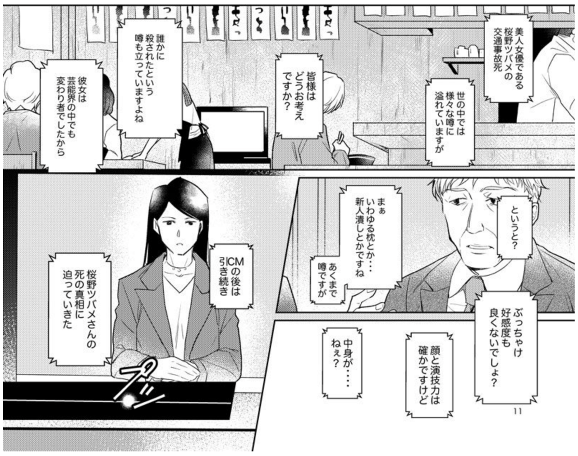

美人女優である桜野ツバメの交通事故死
世の中では様々な噂に溢れていますが
皆様はどうお考えですか？
誰かに殺されたという噂も立っていますよね
彼女は芸能界の中でも変わり者でしたから
というと？
まぁいわゆる枕とか…新人潰しとかですね
あくまで噂ですが
CMの後は引き続き
桜野ツバメさんの死の真相に迫っていきた
プツッ
ぶっちゃけ好感度も良くないでしょ？
顔と演技力は確かですけど
中身が……ねぇ？
11

Figure 7: Frames preceding the one used for our example in figure 1 used for the ablation study. ©Kira Ito

The prompt used for COD is shown in figure 6.

C Details of MQM Human Evaluation

The goal of using MQM is to produce a method of human evaluation that is consistent, efficient, and sufficiently granular. We use MQM with a manga-specific list of issue types, covering different error types:

- Fluency
  - Punctuation
  - Orthography (spelling, punctuation)
  - Grammar (is it ungrammatical or not)
- Accuracy
  - Addition or omission
  - Mistranslation
  - Untranslated text
- Proper Nouns / Terminology
  - Orthography
  - Failed to recognize as proper noun
- Style
  - Formality
  - Awkward
  - Boring
  - Tone (emotional tone is miscalibrated)
- Other
  - Other

These types are not used in actual score computation, but they str useful for helping us understand the problems of a given piece of translated text. Each error type is assigned a severity level by the evaluating translator, ranging from minor to critical, depending on the impact of the issue on overall quality. MQM provides a scoring system that allows for the calculation of overall quality scores based on the number of identified issues and their severity levels. The MQM score is computed using the following equation

$$S = 1 - \frac{5 \times C_{Min} + 10 \times C_{Maj} + 25 \times C_{Crit}}{\text{Total Word Count}} \tag{1}$$

where C_{Min}, C_{Maj}, and C_{Crit} are the number of errors with a severity of minor, major, and critical, respectively. The evaluating translator decides for each error what the most appropriate severity would be.

D Full Prompts

This section includes all the prompts used as part of our experiments. Only the JA-EN prompts are shown, as the only difference between them and the JA-PL prompts is that the target language needs to be explicitly specified in the prompt 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 8 to figure 16 show the prompts used for all of our approaches.

E Visual Feature Ablation Study

To better understand the role of visual features in improving translation accuracy, we conduct an ablation study. Specifically, we systematically obscure parts of the final frame (shown in figure 7) preceding the one used for our example in figure 1 and measure the impact on performance for the corresponding translation. We mask the television, including its “off” sound symbol, the presenter, the surrounding background, and unrelated areas including the counter the TV is standing on. When the key region, i.e., the border of the TV and its “off” symbol, is obscured, the translation accuracy for that particular sentence using PBP-VIS decreases significantly compared to when it is visible – falling to performance comparable to PBP. We do not observe this drop in accuracy for other masked regions. We observe the same behavior for LBL-VIS and LBL. This suggests that the visual feature of the television, along with its symbolic representation of it being switched off, plays a crucial role in the model’s ability to correctly interpret the context for this example.

```
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: {self.line}
Your task is to translate the line to {self.lang}.
Return the translated line in {self.lang} in square brackets [].
Example: {self.jp_example} Return: [{self.lang_example}]
```

Figure 8: Prompt used for LBL approach.

```
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: {self.line}.
Provide the translated lines in square brackets [], without any additional words or characters.
    Provide only one translation for each line.
Example: {self.jp_example} Return: [{self.lang_example}]
```

Figure 9: Prompt used for PBP approach.

```
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: {self.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.
Example: {self.jp_example} Return: [{self.lang_example}]({self.img_explanation_example}).
```

Figure 10: Prompt used for LBL-VIS approach.

```
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:
{self.page}

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: {self.jp_example} Return: [{self.lang_example}]({self.img_explanation_example}).
```

Figure 11: Prompt used for PBP-VIS approach.

```
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:
{self.page}
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: {self.jp_example} Return: Translation 1: [{self.lang_example}]({self.
    img_explanation_example}).
```

Figure 12: Prompt used for PBP-VIS-NUM approach.

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

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:
{self.page}

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: {self.jp_example}

Return:
(
   \”story_jp\”: \”{self.jp_story}\”,
   \”story_en\”: \”{self.lang_story}\”,
   \”lines\”: [
   (
      \”line\”: \”{self.jp_example}\”,
      \”speaker\”: \”{self.lang_speaker}\”,
      \”situation\”: \”{self.lang_situation}\”,
      \”translation\”: \”{self.lang_example}\”,
      \”explanation\”: \”{self.lang_explanation}\”,
   ),
   ]
)
```

Figure 13: Prompt used for VBP-VIS-COD approach.

```
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:
{self.page}

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: {self.jp_example}

Page 2:
Line 1: {self.jp_example2}

Return:
(
   \"pages\": [
   [
   (
      \"line\": \"{self.jp_example}\",
      \"translation\": \"{self.lang_example}\",
      \"reasoning\": \"{self.lang_resoning}\",
   ),
   ],
   [
   (
      \"line\": \"{self.jp_example2}\",
      \"translation\": \"{self.lang_example2}\",
      \"reasoning\": \"{self.lang_resoning2}\",
   ),
   ],
   ]
)
```

Figure 14: Prompt used for VBP-VIS-3P approach.

```
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:
{self.pages}

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

Your task is to translate the lines from the next untranslated page - page {self.curr_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: {self.jp_examplee}

Page 2:
Line 1: {self.jp_example2}

Page 3:
Line 1: {self.jp_example3}

Page 1:
Translation 1: {self.lang_example}

Return:
(
   \”lines\”: [
   (
      \”line\”: \”{self.jp_example2}\”,
      \”translation\”: \”{self.lang_example2}\”,
      \”reasoning\”: \”{self.lang_resoning2}.\”,
   ),
   ]
)
```

Figure 15: Prompt used for VBP-VIS-ALL approach.

```
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 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: {self.jp_example}

Page 2:
Line 1: {self.jp_example2}

Return:
(
   \"pages\": [
   [\"{self.lang_example}\"],
   [\"{self.lang_example2}\"],
   ]
)
```

Figure 16: Prompt used for VBV-VIS approach.