Domain Adaptation of Image Encoder for Multimodal Manga Translation

Kota Manabe
Ehime University

Tomoyuki Kajiwara
Ehime University

Takashi Ninomiya
Ehime University

The University of Osaka
{manabe@ai.cs., kajiwara@cs., ninomiya.takashi.mk@} ehime-u.ac.jp

Isao Goto
Ehime University
goto.isao.fn@ehime-u.ac.jp

Shonosuke Ishiwatari
Mantra Inc.
ishiwatari@mantra.co.jp

Hiroshi Noji
Mantra Inc.
noji@mantra.co.jp

Abstract

The objective of this paper is to enhance machine translation for manga (Japanese comics) by developing and employing an image encoder that is capable of more accurately comprehending its visual context. Conventional manga machine translation systems have faced the challenge of lacking sufficient manga comprehension capabilities when utilizing image information. To address this issue, we propose a domain-adapted image encoder training method for manga. The proposed method involves training encoders to acquire visual features that consider the structural and sequential characteristics of the manga. This approach draws upon a technique that has proven to be highly effective in training language models. The image encoders trained by the proposed methods are used as visual processors in a multimodal machine translation model, and they are evaluated in a Japanese-English translation task. The experimental results demonstrate that the proposed method enhances the performance metrics for translation evaluation, such as BLEU and xCOMET, in comparison to the conventional method.

1 Introduction

Manga, a unique form of expression combining illustrations and text, also known as Japanese comics, is an important part of Japanese culture that has gained popularity worldwide. Although the global demand for manga is growing, manual translation is time-consuming and costly. To address these issues, research on machine translation of manga (Hinami et al., 2021; Lippmann et al., 2025), a technology that automatically translates the text in speech balloons within panels from one language to another, has been studied. However, when only the text in speech balloons is translated, accurate translation is often difficult due to the omission of subjects or lack of context. To solve this problem,

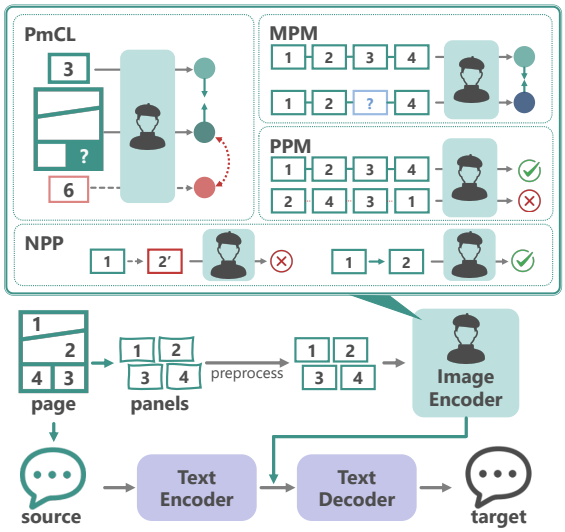


Figure 1: Overview of our multimodal manga translation model. We train domain adaptation of our image encoder to manga using four proposed methods.

context-aware translation (Tiedemann and Scherrer, 2017; Hinami et al., 2021) was developed. In this method, the context is supplemented by adding the immediately preceding text. However, although images contain information that cannot be conveyed by text alone, such as character expressions, backgrounds, and story development, this visual information has not yet been effectively utilized. In response, multimodal machine translation (MMT) (Huang et al., 2016; Delbrouck and Dupont, 2017b,a; Calixto and Liu, 2017; Yao and Wan, 2020; Yin et al., 2020; Sulubacak et al., 2020; Zhang et al., 2020; Wu et al., 2021; Cheng et al., 2024), which uses images as visual information, is being developed. However, most existing methods of utilizing visual information are limited to using comic images' content as tags or captions (Hinami et al., 2021; Saito and Matsui, 2015). These approaches depend on the accuracy of image recognition and fail to utilize information that considers the manga's unique flow and structure. Addition-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 17–26

ally, even when image features are used (Lippmann et al., 2025), general-purpose image encoders are employed, and there have been cases where performance did not improve as expected. One possible reason for this is that conventional image encoders may not be able to capture the information derived from manga images accurately. In other words, conventional image encoders may not deeply understand manga.

To address the challenge that existing manga translation models do not fully utilize manga image information, we propose a training method to acquire visual latent representations that consider the unique structure and flow of manga, based on techniques that have brought significant benefits to language models. Figure 1 illustrates the overview of our method. The objective of this research is to enhance the precision of machine translation for manga by equipping the encoder with the ability to more accurately comprehend manga, and it puts forward a new training approach. Specifically, we apply four methods that have greatly benefited the field of natural language processing to manga image training: contrastive learning (CL) (Chen et al., 2020), masked language modeling (MLM), next sentence prediction (NSP) (Devlin et al., 2019), and permutation language modeling (PLM) (Yang et al., 2019). We propose the applied training methods as Panel matching CL (PmCL), Masked Panel Modeling (MPM), Permutation Panel Modeling (PPM), and Next Panel Prediction (NPP), respectively.

Our experimental results demonstrate that the proposed methods significantly improve automatic translation evaluation metrics such as BLEU and xCOMET compared to conventional baselines. Additionally, our analysis confirms that providing visual information improves translation accuracy and shows that a domain-adapted image encoder for manga is necessary for accurate translation.

2 Related Work

This section offers a comprehensive review of recent developments in manga translation and discusses image encoders that have become widely used.

2.1 Manga Translation

In previous research on manga translation, accuracy has been enhanced by incorporating contextual text information (Tiedemann and Scherrer, 2017; Hinami et al., 2021) and metadata such as author and genre (Kaino et al., 2024). However, these studies generally do not leverage manga image information, indicating room for further improvement. Similarly, methods that combine object detection and OCR (Narasimhan and Singh, 2025) execute translation pipelines from speech bubble detection to integration of the translated text into the image, but they do not utilize image information during the translation process itself.

Attempts to integrate visual tags from images using Illustration2vec (Saito and Matsui, 2015) have been limited by existing image features' ability to adequately represent manga elements, leading to instances of incorrect tag assignments. Although translation with multimodal large language models (MLLMs) (Lippmann et al., 2025) is advancing, there are concerns about inference costs and the potential for further improvement in the manga-specific specialization of current image encoders. Consequently, methodologies for imparting text understanding capabilities that account for manga's unique structure and flow to an image encoder remain insufficiently investigated.

3 Preliminary: Effective Training Methods for NLP

This section delineates CL, a widely utilized approach in image training, and training methods that have made substantial contributions to natural language processing.

3.1 Contrastive Learning

CL (Chen et al., 2020) is a self-supervised training method that learns highly generalizable features, even from unlabeled data. It accomplishes this by minimizing the distance between similar features and maximizing it between dissimilar features. CL has gained considerable attention, particularly in the contexts of unsupervised and self-supervised learning, and its effectiveness is widely recognized.

A common strategy is to apply data augmentation to the input data. For image data, augmentations such as cropping and flipping are applied to generate images from different perspectives. If the original image is considered the anchor, the data-augmented images are used as positive examples, while those generated from other data within the batch are used as negative examples. The widely used InfoNCE loss (van den Oord et al., 2019) is

defined as follows:

$$\mathcal{L}_{\text{InfoNCE}}(h_i, h_j) = -\log \frac{\exp(\text{sim}(h_i, h_j)/\tau)}{\sum_{k \in N} \exp(\text{sim}(h_i, h_k)/\tau)} \quad (1)$$

Here, h_i and h_j denote the anchor feature and the positive example feature, respectively. The sim function is employed to calculate the similarity between two features, with the cosine similarity method being the primary approach. N samples contain one positive example and $N-1$ negative samples, and it is designed to maximize the similarity of positive example pairs within a batch while minimizing the similarity of negative example pairs. The parameter τ is a temperature that controls the learning process.

This results in a model capable of obtaining features that are usable for a variety of tasks. In this research, we will also apply this as a training method by using positive and negative examples suitable for manga.

3.2 Masked Language Modeling

Unlike conventional language models, which can only utilize context unidirectionally, MLM (Devlin et al., 2019) allows for training that considers bidirectional context. In MLM, a portion of the input text is masked, and the model predicts the masked words. The model is trained to extract information from the surrounding context of masked tokens and infer the appropriate words for them. The objective function for MLM is:

$$\mathcal{L}_{\text{MLM}}(\theta) = -\sum_{t=1}^{T} m_t \log p_\theta(x_t|\hat{\mathbf{x}}) \quad (2)$$

$$p_\theta(x_t|\hat{\mathbf{x}}) = \frac{\exp(h_\theta(\hat{\mathbf{x}})_t^\top e(x_t))}{\sum_{x'} \exp(h_\theta(\hat{\mathbf{x}})_t^\top e(x'))} \quad (3)$$

In this equation, the value of $m_t = 1$ signifies that token x_t is masked, while $h_\theta(\hat{\mathbf{x}})_t$ denotes the features extracted from the encoder. Additionally, $e(x_t)$ denotes the embedding of x_t. The parameter θ is optimized to enhance the predicted probability of the correct word by calculating the cross-entropy loss using the predicted probability of occurrence for the mask. This enables the model to process context-dependent word meanings, thereby facilitating the generation of more nuanced representations for each word.

3.3 Next Sentence Prediction

NSP (Devlin et al., 2019) is a task that trains a model to understand the relationship between two sentences by predicting whether they are consecutive. Specifically, the training data consists of pairs of either two consecutive sentences or two randomly extracted, unrelated sentences.

This objective entails not solely the consideration of word-level relationships but also the examination of semantic and logical connections between sentences. Achieving the objective enables the system to function effectively even in cases where it is necessary to span multiple sentences, such as in question answering and summarization.

3.4 Permutation Language Modeling

To retain the advantages of traditional autoregressive language models, which are capable of learning the dependencies between words to be predicted, while also gaining the ability to handle forward and backward information simultaneously, a unique pre-training task called PLM (Yang et al., 2019) was developed.

In PLM, the tokens in the input text are randomly permuted, and some are selected as the target for prediction. In this process, the remaining tokens are employed as conditions while preserving the autoregressive nature rather than masking the tokens. The objective function is as follows:

$$\mathcal{L}_{\text{PLM}}(\theta) = -\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T} \left[\sum_{t=1}^{T} \log p_\theta(x_{z_t}|\mathbf{x}_{\mathbf{z}<t}) \right] \quad (4)$$

Here, $\mathcal{Z}_T$ denotes the set of all permutations of length T, and z represents one such permutation. Note that $p_\theta(x_{z_t}|\mathbf{x}_{\mathbf{z}<t})$ is calculated in the same way as in Equation (3) for the t-th token x_{z_t} in the permutation z. This method allows the model to learn not only the context of the sentence but also the potential dependencies between all tokens.

4 Proposed Method

In this section, we propose a framework for constructing an image encoder capable of comprehending the visual information inherent in manga. The training of this manga-specialized image encoder consists of four adapted training methods: Panel matching CL (PmCL), Masked Panel Modeling (MPM), Next Panel Prediction (NPP), and Permutation Panel Modeling (PPM). As Figure 2 illustrates, the training process based on CL is depicted, while Figure 3 provides a comprehensive representation of the training procedure based on MLM, NSP, and PLM.

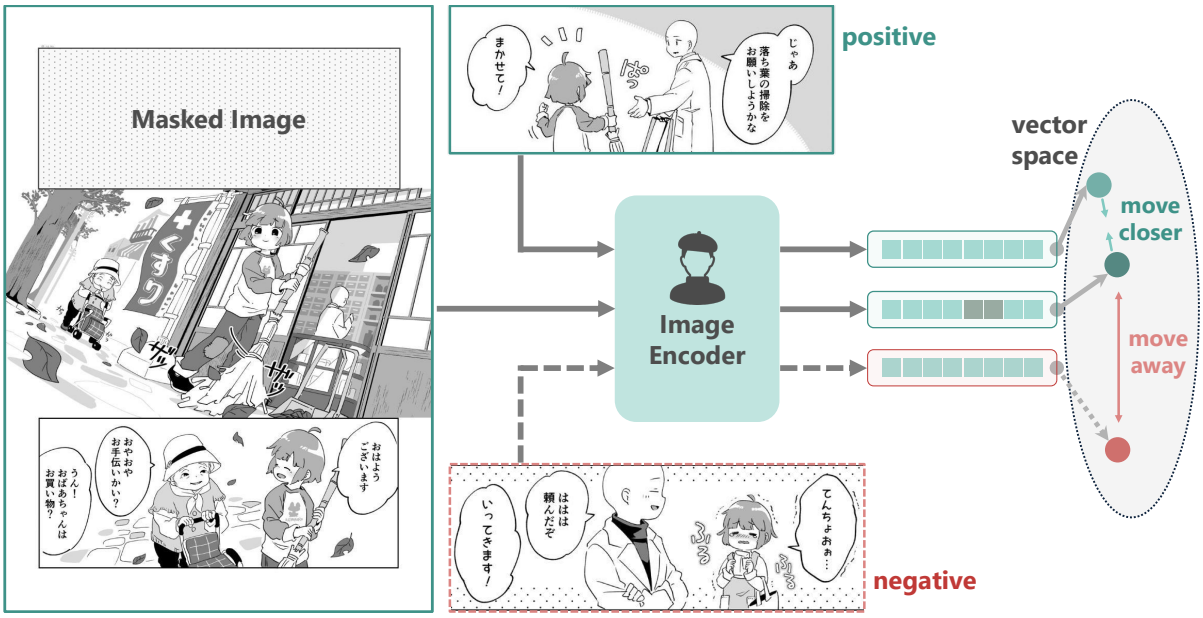


Figure 2: Training methods based on CL (PmCL). Adapted from OpenMantra dataset, licensed under CC BY-NC 4.0 © Nako Nameko

4.1 PmCL: Panel matching CL

The utilization of CL results in an enhancement of the similarity between the masked panels and the correct panels, while concurrently reducing similarity with other panels. Consequently, the model is trained to predict panel images that are appropriate for the masked regions.

As Figure 2 illustrates, we prepare paired data consisting of masked comic page images and correct panel images corresponding to the masked areas as input. We create these masked images and panel images using the coordinate information for each panel. These are then fed into the image encoder to obtain feature vectors (E_{mask} and E_{pos}). We also encode other panels (negative examples) within the same batch to get their feature vectors (E_{neg}). The encoder is trained using the in-batch negative example method to maximize the cosine similarity between E_{mask} and E_{pos} and minimize it between E_{mask} and E_{neg}. It is important to note that the dataset for this training method is pre-created using masked images and panel images, derived from the coordinate information of the panel images.

4.2 MPM: Masked Panel Modeling

Regarding MPM, specific panels within a page are masked, and their content is predicted based on the context of other panels within the page. As Figure 3 demonstrates, the system receives multiple panel images of the entire manga page, with specific panels intentionally masked. Subsequent to the embedding of all multiple-panel images into features, the areas corresponding to the panel images are replaced with masks at a specific ratio. Subsequently, the multiple panel images with masks (I_{masked}) and the panel images of the entire page (I_{original}) are processed by the encoder. Finally, the cosine similarity between the masked and unmasked panels is calculated for the masked areas, and the model is trained to maximize this similarity. Consequently, the encoder assimilates the interrelationships among panels within a page, thereby gaining the capability to supplement absent panel content.

4.3 NPP: Next Panel Prediction

Based on NSP, the model learns to predict whether the panel following a specific panel is an actual consecutive panel or a non-consecutive panel randomly sampled from another page. In Figure 3, $I_{\text{pos\_pair}}$ represents two consecutive panel images, while $I_{\text{neg\_pair}}$ indicates two non-consecutive panel images. Consecutive panels are taken from the same page, while panel images from different pages are used as negative examples. The preprocessed frames are integrated and passed through an image encoder to generate feature vectors. These feature vectors are then entered into a binary classifier, specifically a multi-layer perceptron (MLP), to predict the consecutiveness of the two panels.

This capability allows the encoder to discern the logical flow and continuity of manga panels.

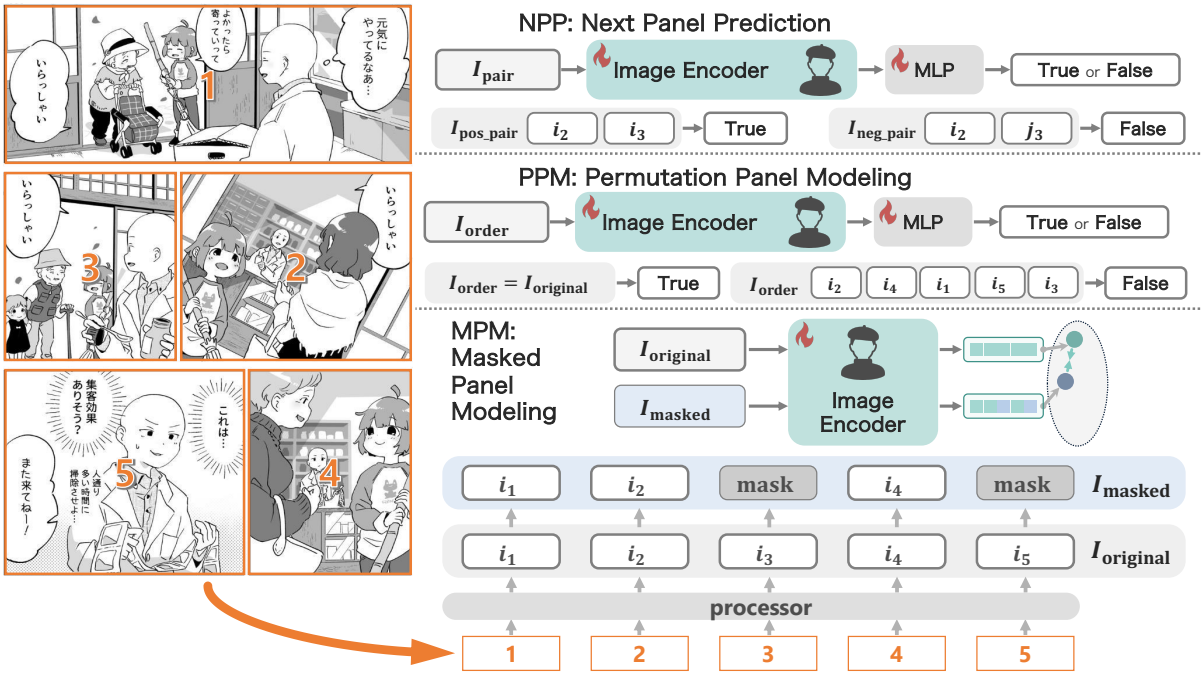


Figure 3: Training methods based on MLM, NSP, PLM (MPM, NPP, PPM). Adapted from OpenMatra dataset, licensed under CC BY-NC 4.0 © Nako Nameko

4.4 PPM: Permutation Panel Modeling

Applying the concept of focusing on permutations in PLM to manga images, the model learns to predict the correct order when the panels within a manga page are shuffled. Figure 3 shows that the input consists of randomly shuffled panels (I_{shuffled}) and panels in their original order (I_{original}) within a manga page. Feature extraction is performed by preprocessing each panel and feeding the combined features to the image encoder. The extracted features are subsequently processed by an MLP to predict the order of the panels. The model is trained with the shuffled input labeled as "False" and the original input labeled as "True."

5 Experiments

To verify the effectiveness of the proposed method, an evaluation experiment was conducted on a Japanese-English translation task using the Manga Corpus constructed in previous research (Hinami et al., 2021).

5.1 Data

During the training phase of the image encoder, we employed page images sourced from the Manga Corpus (Hinami et al., 2021). For validation and evaluation data, two volumes from the latest release of each title were pre-extracted, with the remaining volumes used as training data. From the extracted data, 1,000 image and text pairs were randomly sampled for validation and evaluation.

Since image and text pairs are required for translation, image-only data were removed before MMT training. By using pre-recorded panel coordinate information, masked image and corresponding panel image pairs were created, in addition to all panel images.

5.2 Setup

For image preprocessing, we utilized an ImageProcessor from the Hugging Face Transformers library (Wolf et al., 2020). This involved resizing all input images to a uniform dimension. The encoder models employed in this study included ViT[1], pre-trained on ImageNet (Deng et al., 2009), and CLIP[2], trained on large-scale internet data. These models are generic image models, not specialized for manga. Following the training methods outlined in Section 4 (PmCL, MPM, NPP, and PPM), each model was trained with these individual methods, as well as with a combination of all methods. Also, for the MPM, 0.65 was adopted for the masking ratio. For the model architecture, the MLP layers used in NPP and PPM consisted of two linear layers with ReLU activation functions (Nair

[1] https://huggingface.co/google/vit-base-patch16-224-in21k

[2] https://huggingface.co/openai/clip-vit-base-patch16

and Hinton, 2010) and a fully connected layer. Received features were linearly transformed to the hidden dimension size of the language model, then progressively converted down to 256 dimensions while applying the ReLU function, and finally transformed to the number of labels. Contrastive loss was used for PmCL and MPM, while cross-entropy loss was employed for NPP and PPM, with models saved every 500 steps. Training was terminated when the loss did not improve for 10 consecutive model updates.

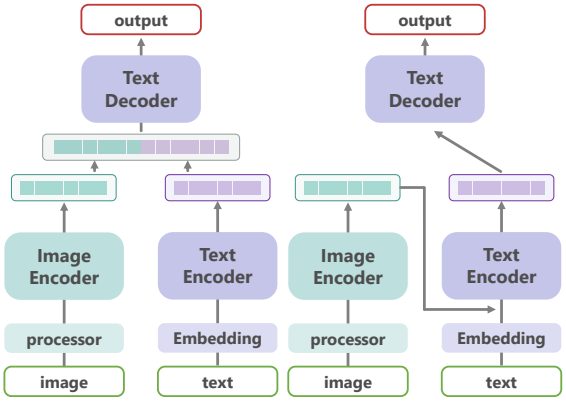
Considering the impact of model size, we adopted mT5 (Xue et al., 2021) base[3] and large[4], which are pre-trained multilingual models, as the base models for the MMT models, i.e., we used them as the text processing components. For the visual processing component, we employed the aforementioned image encoders.

During MMT training, we used the tokenizer provided by Hugging Face for text preprocessing, employing tokenization suitable for mT5. For Japanese-to-English translation, "translate Japanese to English: " was added to the source language text. Image preprocessing was similarly used with its dedicated processor. For data, Japanese text and the comic panel image containing the text were prepared as input, with English text as the output. Only the image encoder's parameters were kept fixed, and all other parameters were trained for the MMT model. Cross-entropy loss was used, and training was stopped if the loss value didn't decrease for five consecutive MMT model updates (checked every 1,000 steps). We used a batch size of 32 and followed the default settings of the Hugging Face trainer for the other training parameters.

5.3 Modality Fusion

As illustrated in Figure 4, two fusion methods were employed to combine visual and linguistic information. A straightforward concatenating approach (Danapal et al., 2020; Steinbaeck et al., 2018), termed parallel encoding, involved concatenating the feature vectors output by the image encoder and text encoder before passing them to the decoder. When combining in parallel, an attention mask for the visual features was pseudo-created and combined with the attention output by the text encoder to direct attention.

[3] https://huggingface.co/google/mt5-base
[4] https://huggingface.co/google/mt5-large



(a) parallel (b) serial

Figure 4: Fusion methods for vision and language

Alternatively, a method termed serial encoding involved combining the features obtained from the image encoder with the text embeddings (Scalzo et al., 2008; Li and Wu, 2019), thereby providing the text encoder with an input that incorporates visual information. For text embedding, the embedding layer in the encoder was used to obtain embeddings for the input sentence, and visual features were added to these embeddings. In this case, an attention mask was similarly pseudo-created and utilized for attention. These two fusion methods were leveraged in the MMT model for modality fusion.

5.4 Baseline and Evaluation Methodology

We compare our proposed method with two approaches: text-only translation models that do not utilize any image information and MMT models combining a pre-trained general-purpose image encoder with a text encoder. In both baseline cases, mT5 was fine-tuned on the training data.

Regarding our proposed method, we evaluate models that incorporate image encoders trained using the four aforementioned training methods (PmCL, MPM, NPP, and PPM) for the visual processing part of the MMT model. This includes both models trained with individual methods and those trained by combining all four. Furthermore, we comprehensively evaluated the combinations of fusion methods presented in Section 5.3. By comparing these translation results, we quantitatively demonstrate the impact of the proposed method and the fusion methods on translation performance. For evaluation, we used sacreBLEU[5] (Post, 2018) as an automatic evaluation metric widely used in

[5] https://github.com/mjpost/sacrebleu

mT5-base					mT5-large				
Enc.	Fusion	Variant	BLEU	xCOMET	Enc.	Fusion	Variant	BLEU	xCOMET
–	–	text-only	10.47	0.681	–	–	text-only	13.08	0.734
ViT	parallel	baseline	8.06	0.612	ViT	parallel	baseline	14.05	0.741
		pmcl	**11.36**	**0.692**			pmcl	13.05	0.737
		mpm	**11.21**	**0.696**			mpm	13.78	0.739
		npp	**11.82**	**0.705**			npp	**14.82**	**0.753**
		ppm	**11.73**	**0.723**			ppm	13.61	**0.745**
		all	**13.82**	**0.745**			all	**15.36**	**0.777**
	serial	baseline	12.09	0.710		serial	baseline	15.98	0.779
		pmcl	11.32	0.697			pmcl	**16.21**	0.778
		mpm	11.54	0.697			mpm	**16.44**	**0.780**
		npp	**13.78**	**0.726**			npp	15.53	0.773
		ppm	**14.64**	**0.746**			ppm	**16.15**	0.778
		all	**14.94**	**0.761**			all	**16.40**	**0.782**
CLIP	parallel	baseline	13.23	0.732	CLIP	parallel	baseline	14.92	0.755
		pmcl	12.30	0.727			pmcl	13.95	0.745
		mpm	**13.58**	0.731			mpm	14.01	0.745
		npp	**13.30**	0.727			npp	**14.99**	**0.762**
		ppm	11.54	0.696			ppm	14.08	0.743
		all	**13.65**	**0.739**			all	**15.43**	**0.785**
	serial	baseline	12.46	0.720		serial	baseline	15.91	0.779
		pmcl	**12.88**	**0.724**			pmcl	15.11	0.772
		mpm	**14.82**	**0.742**			mpm	14.89	0.768
		npp	**12.62**	0.711			npp	15.84	0.771
		ppm	**12.53**	0.715			ppm	15.86	0.776
		all	**14.23**	**0.741**			all	**16.18**	**0.787**

Table 1: Results in Japanese-English manga translation. The highest performance for each language model is underlined, and methods outperforming baselines are bolded. SacreBLEU is used for BLEU scores. For each encoder, the variants "pmcl", "mpm", "npp", "ppm", and "all" are compared against the corresponding "baseline".

translation tasks, which mainly assesses surface-token agreement. Following sacrebleu's default settings, the tokenizer used 13a and considered N-gram precision up to a maximum of 4-grams. Additionally, we utilized xCOMET[6] (Guerreiro et al., 2024), which considers semantic meaning, for a multifaceted evaluation.

5.5 Results

Table 1 presents the evaluation results for the translation task across different language models and image encoders. One of the proposed methods, which integrates and trains the four training frameworks, is denoted by the suffix "all." The experimental findings demonstrated that the MMT model incorporating the image encoder trained with the integrated proposed approach consistently exhibited significant enhancements across both metrics, irrespective of model size or fusion method, when compared to both the text-only baseline and the multimodal baseline using a general-purpose image encoder. Specifically, for mT5-base with ViT under serial fusion, our integrated method (all) improved BLEU from 12.09 to 14.94 and xCOMET from 0.710 to 0.761, compared to the corresponding baseline. Furthermore, the consistent superiority of the proposed method over MMT models using general-purpose image encoders across all combinations also demonstrated the effectiveness of the domain-adapted image encoder.

6 Analysis

A detailed analysis will be conducted from two perspectives, as indicated by the experimental results. The first perspective will involve an analysis of translation performance concerning the proposed methods and fusion techniques. The second perspective will be a case analysis of the results.

[6] https://huggingface.co/Unbabel/XCOMET-XL

	mt5-base	
Variant	BLEU	xCOMET
baseline	11.46	0.694
pmcl	11.96	0.710
ppm	12.61	0.720
mpm	12.79	0.717
npp	12.88	0.717
all	14.16	0.746

Table 2: Average MMT evaluation score for each training method of image encoders. Here, "baseline" denotes the MMT setting that uses a general-purpose image encoder without domain adaptation.

source text	疲れてたのかもな
reference	Maybe he was tired.
text-only	Maybe she was tired or something.
baseline	Maybe she was tired.
all	Maybe he was tired.

Figure 5: Translation examples with the corresponding manga panels. The "baseline" and "all" settings use ViT as the image encoder. Adapted from OpenMatra dataset, licensed under CC BY-NC 4.0 © Nako Nameko

6.1 Performance of Each Translation Method

We analyze the impact of each proposed training method and different feature fusion methods on translation performance. Additionally, performance disparities are examined based on the utilized image encoder model.

Table 1 also illustrates the impact of training methods and fusion approaches on translation quality. Most methods incorporating image information showed improvements in both BLEU and xCOMET compared to the text-only baseline. When comparing our MMT models applying individual proposed methods to pre-trained encoders and MMT models using pre-trained encoders, the mt5-base variant demonstrated score improvements in 12 out of 16 methods. Meanwhile, the mt5-large variant surpassed in 5 methods. Furthermore, table 2 shows that when comparing the proposed methods with pre-trained general-purpose encoders, consistent score improvements were observed across all methods for mT5-base. Among them, the -all method, which combines the four techniques, is found to be the best. This indicates the importance of combining the four proposed training methods. Additionally, greater performance improvements were observed with smaller model sizes.

Comparing the averages across fusion methods, serial encoding consistently outperformed parallel encoding, regardless of model size. Particularly in BLEU, an average performance difference of more than 1 point was observed. This indicates that for MMT models, early fusion of visual features allows for more effective utilization of that information. Regarding image encoders, ViT demonstrated a slight performance advantage over CLIP in BLEU, while a significant performance difference due to the combination of methods was prominently observed in both evaluation metrics.

6.2 Case Analysis

We analyze the impact of our proposed method by comparing translation examples generated by our model against those generated by the conventional models. Specifically, we compare translation examples from the mt5-large model employing vit-all-serial, as significant performance improvements were observed with this configuration in evaluation metrics. Figure 5 presents actual source language texts and the corresponding outputs from each model. While the conventional translation method mistranslated the gender, the proposed method correctly translated it as "he."

7 Conclusion

This research proposes a training method for constructing an image encoder capable of "understanding" the visual context of manga, aiming to overcome the limitations of existing MMT models in fully leveraging manga image information. In evaluation experiments, integrating a domain-adapted image encoder trained with our proposed method as the visual processing component of an MMT model consistently yielded significant improvements in translation evaluation metrics, compared to both baseline models. We anticipate these findings will both improve manga translation capabilities and foster the global spread of manga culture. Our future work will focus on overcoming challenges to

enhance performance further. Specifically, we intend to explore optimal parameters for each training method, integrate more context-aware approaches, and extend our methodology to MLLMs.

Limitations

In this study, we utilized the relatively lightweight mT5 language model to verify the benefits of image encoders at low cost. On the other hand, translation using MLLMs combined with an image encoder specialized for manga remains a future challenge.

Furthermore, evaluation relies on automated metrics such as BLEU and xCOMET. While these quantitatively indicate translation quality, they struggle to fully capture qualitative aspects like the reproducibility of character speech patterns or contextually appropriate paraphrasing. This limitation also applies to human evaluation, requiring expert assessment to determine whether a translation is optimal after understanding the narrative context.

Acknowledgments

These research results were obtained from the commissioned research (No.22501) by the National Institute of Information and Communications Technology (NICT), Japan.

References

Iacer Calixto and Qun Liu. 2017. Incorporating Global Visual Features into Attention-based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. SoulMix: Enhancing Multimodal Machine Translation with Manifold Mixup. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11283–11294.

Gokulesh Danapal, Giovanni A. Santos, João Paulo C. L. da Costa, Bruno J. G. Praciano, and Gabriel P. M. Pinheiro. 2020. Sensor fusion of camera and LiDAR raw data for vehicle detection. In *2020 Workshop on Communication Networks and Power Systems*, pages 1–6.

Jean-Benoit Delbrouck and Stephane Dupont. 2017a. Multimodal Compact Bilinear Pooling for Multimodal Neural Machine Translation. *arXiv:1703.08084*.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017b. Modulating and attending the source image during encoding improves Multimodal Translation. *arXiv:1712.03449*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, pages 979–995.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12998–13008.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 639–645.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 17337–17342.

Hui Li and Xiao-Jun Wu. 2019. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Transactions on Image Processing*, 28(5):2614–2623.

Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. 2025. Context-Informed Machine Translation of Manga using Multimodal Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3444–3464.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 807–814.

Nithyasri Narasimhan and Sagarika Singh. 2025. Crossing Language Borders: A Pipeline for Indonesian Manhwa Translation. *arXiv:2501.01629*.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191.

Masaki Saito and Yusuke Matsui. 2015. Illustration2Vec: a semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*.

Fabien Scalzo, George Bebis, Mircea Nicolescu, Leandro Loss, and Alireza Tavakkoli. 2008. Feature Fusion Hierarchies for gender classification. In *2008 19th International Conference on Pattern Recognition*, pages 1–4.

Josef Steinbaeck, Christian Steger, Gerald Holweg, and Norbert Druml. 2018. Design of a Low-Level Radar and Time-of-Flight Sensor Fusion Framework. In *2018 21st Euromicro Conference on Digital System Design*, pages 268–275.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal Machine Translation through Visuals and Speech. *Machine Translation*, pages 97–147.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal Transformer for Multimodal Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural Machine Translation with Universal Visual Representation. In *International Conference on Learning Representations*.