Integrating the BERT-Deformer Architecture to Optimize Multimodal Translation of Japanese Literature

Li Shan[1]
[1]School of Foreign Languages, Dalian Polytechnic University, Dalian 116034, Liaoning, China
Corresponding author: Li Shan (e-mail: sl2026sl@163.com)

ABSTRACT Accurate multimodal translation of Japanese literature requires effective integration of linguistic semantics and visual context while maintaining long-range dependency modeling and cultural fidelity. This study presents a BERT-Deformer architecture that combines a pre-trained Japanese BERT encoder with deformable attention to achieve hierarchical fusion of textual and visual information. The proposed framework employs deep contextual representations for literary texts and dynamically focuses on salient cross-modal features through adaptive reference-point sampling, enabling efficient processing of long sequences and culturally sensitive semantic alignment. Experimental results demonstrate superior performance over conventional Transformer-based approaches, achieving improvements in BLEU-4, METEOR, and BERTScore while exhibiting enhanced robustness for long-text translation and more accurate handling of culturally loaded expressions. Attention analysis further confirms that the model effectively exploits visual cues to strengthen semantic disambiguation and cross-modal consistency. Beyond literary translation, the proposed architecture provides a scalable multimodal information processing paradigm with potential relevance to semantic communication and intelligent electromagnetic information systems, where reliable cross-modal representation and context-aware information transmission are essential for high-fidelity knowledge delivery.

INDEX TERMS multimodal translation, semantic communication, deformable attention, cross-modal fusion, japanese literature

I. INTRODUCTION

As a vital component of the world cultural heritage, Japanese literature translation is of considerable importance for advancing cross-cultural communication and academic study. As multimodal data become increasingly popular, the desire for translation that combines text and visual information is growing. In this study, the visual data are specifically curated from illustrative materials traditionally associated with Japanese literary works, such as ukiyo-e prints for classical literature and commissioned illustrations for modern texts, ensuring a strong semantic bond between text and image. Established machine translation techniques encounter significant difficulty with such literary texts steeped in culturally loaded words and complex rhetoric. There is an urgent need to explore new technical avenues in order to find balance between semantic fidelity and artistic communication [1-3]. The current problematic issues in multimodal translation of Japanese literature revolve around three main facets. First, literary texts are heavily laden with metaphors, idioms, and historical allusion, so models must have a level of semantic comprehension and meaning that goes beyond the literal meaning [4,5]. Second, long paragraphs create a challenge to the long-range dependency modeling capability of existing neural machine translation architectures, as the computational complexity of attention mechanisms becomes a bottleneck when processing long texts [6,7]. Third, while textual descriptions and the respective visuals have a semantic connection, existing cross-modal alignment methods fail to create fine-grained, non-local associations, leading to poor consistency between the translation outputs and corresponding visual context [8,9].

Prior investigations have sought to address the issues from various aspects, but they have not been completely resolved. Statistical machine translation techniques are limited in their ability to capture the nonlinear aspects that are commonplace in literary language [10,11]. In addition, sequence models that are based on recurrent neural networks are constrained by their capacity for encoding length and the parallelization of the model [12,13]. The standard transformer architecture greatly improves parallel efficiency, but at the same time, the complexity of the self-attention mechanism produces a computation requirement that strongly constrains its practicality for long text translation [14,15]. For multimodal tasks, late-stage fusion approaches do not

account for deep interactions between modalities, while early-stage fusion approaches do not work well because ongoing representation alignment needs to constantly deal with heterogeneity across modalities [16,17]. None of these approaches can simultaneously unify semantic depth, computational efficiency, and performance at the level of modal fusion.

To tackle the challenges of ineffective semantic comprehension, inefficient long-sequence processing, and inadequate cross-modal interaction within current multimodal translation architectures of Japanese literature, this paper presents a fusion architecture based on BERT and a deformable Transformer. In this architecture, a pre-trained Japanese BERT model is used as the text encoder to capture rich linguistic features at both the lexical and syntactic levels. The visual channel employs a Vision Transformer to extract global representations of image patches. In the decoding stage, a multimodal deformable attention decoder is proposed. This decoder dynamically selects key information nodes in the cross-modal sequence through learnable reference points, achieving adaptive focusing and fusion of text and image information with linear complexity. This architecture achieves accurate parsing of complex literary semantics and efficient processing of long documents.

II. BERT-DEFORMER FUSION ARCHITECTURE OPTIMIZES JAPANESE LITERATURE TRANSLATION

A. TEXT AND VISUAL FEATURE ENCODING

The dual-channel encoder constructed in this paper is responsible for transforming heterogeneous Japanese text and associated image inputs into a unified deep feature sequence, providing high-quality semantic and visual representations for subsequent fusion decoding.

Text modality processing employs a pre-trained RoBERTa-japanese-base model as the encoding backbone [18,19]. RoBERTa was chosen over standard BERT due to its improved performance on masked language modeling through dynamic masking and larger batch training, which is particularly beneficial for capturing the nuanced semantics of literary language. While lighter models exist, the superior contextual understanding provided by RoBERTa is deemed essential for the task and justifies its computational cost within the overall architecture. Given a Japanese sentence $S=\{w_1, w_2, ..., w_n\}$, it is first decomposed into a sequence of sub-word units by its dedicated tokenizer, and special classifiers [CLS] and delimiters [SEP] are added. This sub-word sequence is mapped to word embeddings E_{token} and added to position embeddings E_{pos} to form the initial input H_0 of the model. The input is then deep bidirectionally encoded through L Transformer layers of RoBERTa. The computation process of each layer follows the standard Transformer architecture, capturing contextual dependencies through self-attention mechanisms and fully connected feedforward networks [20,21]. Finally, the hidden states of all sub-word positions in the last layer are taken as the text feature sequence $T = \{t_1, t_2, ..., t_n\}$, where $t_i \in \mathbb{R}^{d_t}$, and the vector corresponding to the [CLS] label is used as the global semantic summary of the entire sentence. Visual modality processing is based on the ViT (Vision Transformer) architecture [22,23]. The input literary image I is resized to a fixed size and divided into m non-overlapping flat image patches $\{x_p{}^1, x_p{}^2, ..., x_p{}^m\}$. Each patch is mapped to an embedding space matching the text feature dimensions through a learnable linear projection layer, resulting in a patch embedding e_p. To preserve spatial information, learnable positional embeddings E_{pos} are added to the patch embeddings. Simultaneously, an additional [class] label is preplaced before the patch sequence, and its final state serves as the global representation of the image. The complete embedding sequence $Z_0 = [e_{\text{class}}; e_p{}^1; ...; e_p{}^m] + E_{\text{pos}}$ is fed into the ViT encoder. After processing through multiple Transformer blocks, the output sequence corresponds to the deep visual features of all input blocks and [class] tags, denoted as $V = \{v_{\text{class}}, v_1, v_2, ..., v_m\}$, where $v_i \in \mathbb{R}^{d_t}$, achieving dimensional alignment with the text features. The parallel processing flow of text and visual feature encoding is shown in Figure 1.

Figure 1 illustrates the overall data processing flow of the dual-channel hierarchical feature encoding architecture adopted in this paper. This architecture consists of two parallel channels: text encoding and visual encoding. The text channel first segments and embeds the input Japanese literary text into sub-words, then feeds it into a RoBERTa-Japanese-based Transformer encoder for deep bidirectional semantic encoding, ultimately outputting a text feature sequence rich in contextual information. The visual channel performs block segmentation and linear projection operations on the input image, and then extracts global visual features through a Vision Transformer (ViT) encoder to generate an image block feature sequence. The core modules of both channels achieve deep modeling of their respective modal information through a multi-layer Transformer structure, ultimately outputting dimension-aligned text and visual feature sequences, providing a high-quality semantic and visual representation foundation for subsequent cross-modal fusion in the decoder.

B. MULTIMODAL FUSION DECODING BASED ON DEFORMABLE ATTENTION

This section describes the core design of the decoder, which functions to fuse the text feature sequence $T = \{t_1, t_2,.., t_n\}$ output by the encoder with the visual feature sequence $V = \{v_{\text{class}}, v_1, v_2, ..., v_m\}$ and generate the target language translation autoregressively [24,25]. The decoder L_d consists of stacked layers with identical structures, each containing a standard multi-head self-attention mechanism, a key multimodal deformable attention mechanism, and a feedforward neural network. Given the input vector $h_q{}^{l\text{-}1of}$ the l-th layer decoder, it first passes through a masked multi-head self-attention layer to process the dependencies of its own sequence and outputs

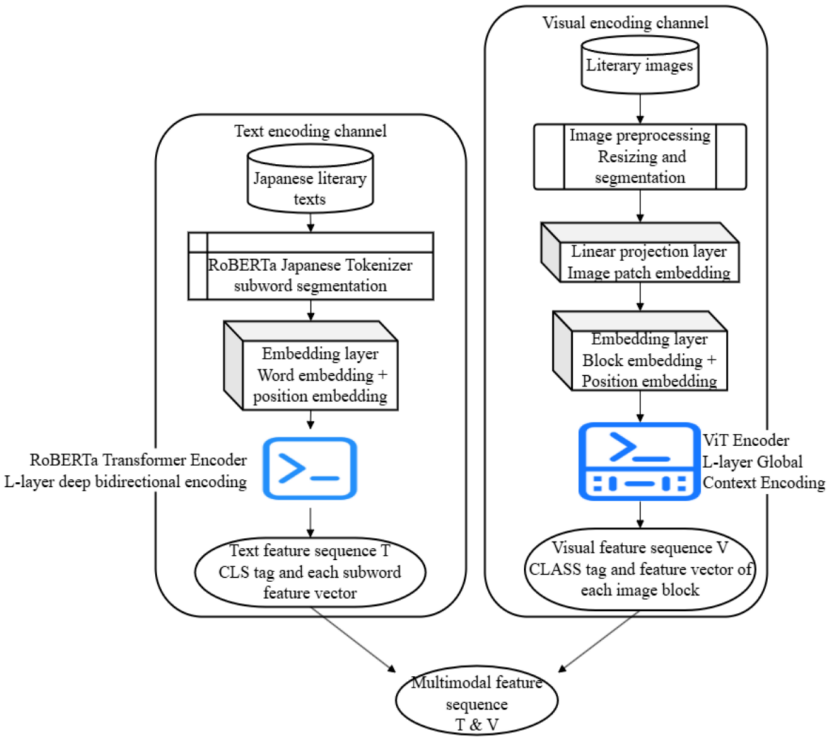

FIGURE 1. Schematic diagram of multimodal feature encoding

an intermediate representation $\hat{h}_q^l$.

This intermediate representation $\hat{h}_q^l$ is then used as input to the core multimodal deformable attention module as a query. This module is designed to efficiently focus key information from a lengthy multimodal context $C = [T; V]$. For each query location q, the mechanism does not calculate its global attention with all keys in the context, but instead predicts the coordinates p_{qk} of a small subset of K important reference points and their corresponding attention weights A_{qk} through a lightweight linear projection layer, as shown in Equation (1):

$$p_{qk}, A_{qk} = \text{Linear}(\hat{h}_q^l) \tag{1}$$

In Equation 1, $k \in \{1,..., K\}$, $K \ll (n + m)$. These dynamically generated reference points p_{qk} are used to sample the K most relevant key-value pairs from the multimodal context C. Finally, the context vector c_q^l of the query location is obtained by weighted summation of the value vectors of these K sampled point, as shown in Equation (2):

$$c_q^l = \sum_{k=1}^{K} A_{qk} \cdot W_V C(p_{qk}) \tag{2}$$

In Equation 2, W_V is the projection matrix of the value vector, and $C(p_{qk})$ represents the feature vector sampled at coordinate p_{qk}. This mechanism reduces the computational complexity from $O((n + m)^2)$ of standard attention to $O(K \cdot (n + m))$, thus enabling efficient processing of long sequence multimodal inputs [26,27]. Beyond efficiency, this dynamic sampling is particularly suited for literary translation because it allows the model to adaptively focus on the most semantically relevant visual patches (e.g., a character's expression, a symbolic object) and textual segments (e.g., a metaphorical phrase) that are sparsely distributed across long narratives. Unlike fixed global attention, which may dilute crucial cues, deformable attention learns to pinpoint these key cross-modal correspondences, directly enhancing the translation of culturally loaded and abstract concepts. The working principle of the dynamic focusing mechanism of the deformable attention module is shown in Figure 2.

Figure 2 illustrates that this mechanism takes the concatenated multimodal context sequence and the current decoder query state as input. The query vector first passes through a linear projection layer, which simultaneously performs two key predictions: first, it generates a set of reference point coordinates that are much smaller than the context length; second, it calculates the attention weight corresponding to each reference point. These dynamically predicted reference points are directly used to accurately sample a few key feature values from the vast multimodal context. Subsequently, the sampled feature values are weighted and summed using the predicted attention weights, and finally fused to generate a context vector output that condenses the most relevant information. The entire process realizes a shift from "global computation" to "dynamic focusing", significantly improving the computational efficiency and relevance of long-sequence multimodal fusion by processing only a small number of key positions.

The context vector c_q^l output by this module is residually

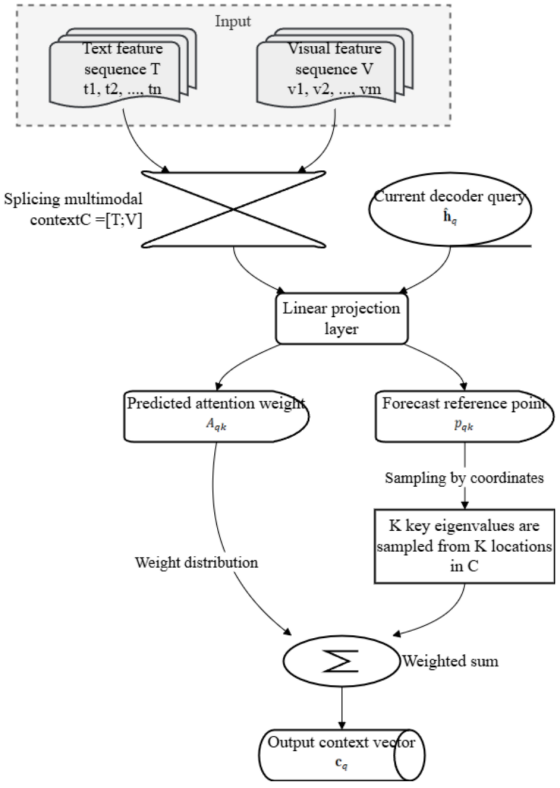

FIGURE 2. Schematic diagram of the working principle of the deformable attention mechanism

connected and layer normalized with the self-attention output $\hat{h}_q^l$, and then fed into the feedforward network for final transformation to obtain the output h_q^l of the decoding layer. This hierarchical processing is performed iteratively in all decoding layers, and finally the probability distribution of the target word is predicted by the output of the last layer through a linear layer and a Softmax function [28,29].

A detailed comparative analysis of the computational properties of deformable attention and standard attention is shown in Table 1.

TABLE 1. Comparison of deformable attention mechanisms and standard attention mechanisms

Comparison dimension	Standard multi-head attention mechanism	Deformable attention mechanism
Calculation principle	Calculate the dot product of the query and all context keys	Query predictions for a small number of reference points, interacting only with sampled keys
Computational complexity	O((n + m)2)	O(K ·(n + m)) K ≪(n + m)
Cross-modal interaction	Implicit, global, computationally intensive	Explicit, local focus, content adaptation
Adaptability to long sequences	Bad, memory and computational costs grow with the square of the sequence length	Strong, linear cost growth, easy to handle long documents
Ability to capture key information	Depends on weight distribution, which may be diluted by irrelevant information	Actively focus on the key positions of prediction, with stronger directionality

Table 1 compares the differences between deformable attention and standard attention mechanisms across five core dimensions: computational principles, complexity, cross-modal interaction, long sequence adaptability, and key information capture. Standard mechanisms, due to their global computational nature, result in quadratic complexity, making them inefficient for handling long multimodal contexts, and their cross-modal interaction is relatively implicit. In contrast, deformable mechanisms dynamically focus on a small number of key reference points through content adaptation, reducing complexity to linear levels. This greatly enhances the processing ability for lengthy documents, while also reaching a high level of precision in obtaining cross-modal key information in a more explicit and targeted way, fundamentally optimizing efficiency and effectiveness in multimodal fusion.

III. EXPERIMENTAL DESIGN

A. DATASET AND EVALUATION METRICS

In order to evaluate the impact of the proposed method on multimodal translation tasks of Japanese literary works, a dataset is created for self-use. The proposed dataset includes a collection of Japanese literary text fragments, alongside high-quality visual illustrations or pictures corresponding to the source text. The target is to have fully developed Chinese versions of professionally translated textual material. The text content is selected from classical and modern Japanese literature, covering various genres such as novels and essays, to ensure diversity of language styles and literary complexity. All image-text pairs have been manually selected

and aligned to ensure a high semantic correlation between the image content and the text description. The dataset is divided into training, validation, and test sets proportionally, and detailed statistical information is shown in Table 2.

TABLE 2. Statistical information of the Japanese literature multimodal translation dataset

Dataset partitioning	Number of picture-text pairs	Average number of characters in Japanese text	Average number of characters in Chinese text	Average image resolution
Training set	45,000	125.4	98.7	768×768
Validation set	5,000	128.1	101.2	772×772
Test set	5,000	123.8	97.5	765×765

This experiment uses a combination of automatic and manual evaluation. Automatic evaluation selects BLEU-4 (Bilingual Evaluation Understudy-4) and METEOR (Metric for Evaluation of Translation with Explicit Ordering) as basic indicators to measure the similarity between the translated text and the reference translation in terms of n-gram matching and synonym reconstruction [30,31]. BLEU-4 index values are expressed in scores, typically ranging from 0 to 1; the closer the value is to 1, the higher the similarity. Given the high requirements for fluency and stylistic consistency in literary translation, this paper additionally applies the BERT-based semantic similarity index BERTScore, calculated as shown in Equation (3). The semantic fidelity is evaluated by comparing the cosine similarity between the generated translation and the reference translation in the deep semantic space [32,33].

$$\text{BERTScore} = \frac{1}{|y|}\sum_{y_i \in y} \max_{\hat{y}_j \in \hat{y}} y_i^T \hat{y}_j \tag{3}$$

In Equation 3, y and $\hat{y}$ represent the BERT context embedding vector sequences of the reference translation and the generated translation, respectively. This measurement of semantic equivalence is useful beyond surface form. Human evaluation investigates fluency, fidelity, and the meaning of culturally loaded words in the translation. Professional translators provide a scoring based on a Likert scale, which adds a valuable human dimension to the automatic evaluation.

B. BASELINE MODEL AND IMPLEMENTATION DETAILS

To comprehensively evaluate the performance of the proposed BERT-Deformer model, four strong baseline models are set up for comparison. The Transformer baseline model adopts the standard encoder-decoder architecture. The Text-Only BERT-Transformer model replaces the encoder of the Transformer with the pre-trained RoBERTa-japanese-base model, while the decoder remains unchanged. This baseline is used to verify the necessity of applying visual modalities. The Multimodal Transformer Late Fusion model, based on the Text-Only model, concatenates the global image features extracted by ViT with the [CLS] tag vector output by the encoder before feeding it into the decoder. This baseline represents a classic late fusion strategy [34]. The Multimodal Transformer Early Fusion model directly concatenates the image block feature sequence with the text sub-word feature sequence at the encoder end to form a long multimodal sequence that is input into the standard Transformer encoder. This baseline is used to verify the effect of coarse-grained modal fusion in the encoding stage.

All models are trained and tested on the same dataset and partition. The implementation of the model and the baseline model in this paper is based on the PyTorch framework and the Transformers library of Hugging Face [35]. The model is trained on 8 NVIDIA A100 GPUs, and the AdamW optimizer is used. The hyperparameter settings are shown in Table 3.

TABLE 3. Model hyperparameter configuration table

Hyperparameters	Configuration values
Optimizer	AdamW
Base learning rate	5e-5
Weight decay	0.01
Training batch size	32
Warm-up steps	1000
Learning rate scheduler	Cosine decay
Discard rate	0.1
Maximum text length	256
Image block size	16×16
Number of training rounds	20
The number of deformable attention reference points k	8

Table 3 shows the model's hyperparameter settings: the learning rate is set to 5e-5; the weight decay coefficient is 0.01; the batch size is uniformly 32. Dynamic learning rate scheduling is used during training, with linear warm-up in the first 1000 steps followed by cosine decay. To prevent overfitting, Dropout with a dropout rate of 0.1 is applied after all fully connected layers and attention layers except for the pre-trained parameters. The maximum length of the text sequence is set to 256 tokens, and the image patch size is set to 16x16. Model training uses the loss on the validation set as the early stopping criterion and lasts for 20 epochs.

IV. RESULTS

A. OVERALL TRANSLATION PERFORMANCE OPTIMIZATION

To systematically evaluate the comprehensive performance and optimization effect of the proposed BERT-Deformer model in the multimodal translation task of Japanese literature, this section quantitatively compares it with a series of strong baseline models on multiple automatic evaluation metrics. Experiments select the Transformer benchmark, the text-only BERT-Transformer, and two multimodal Transformers with different fusion strategies as comparison models. A comprehensive evaluation is conducted from multiple dimensions, including lexical matching and semantic similarity. Rigorous quantitative analysis verifies the effectiveness of the proposed architecture. The comparison results of various metrics are shown in Figure 3.

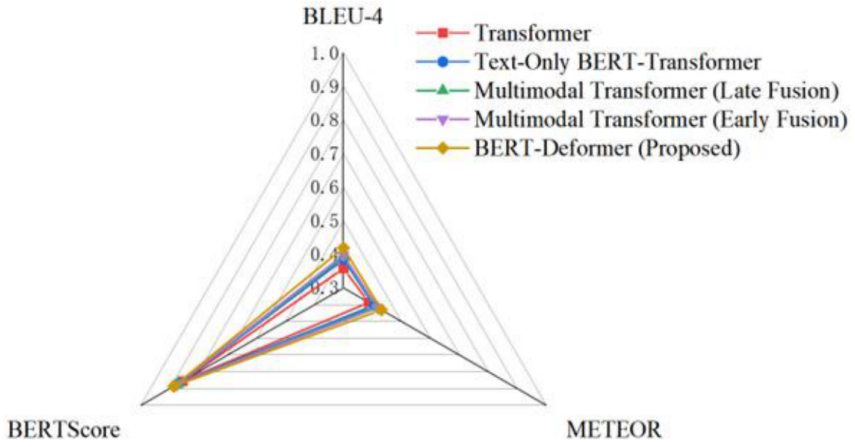

FIGURE 3. Overall performance comparison results of each model

Figure 3 displays the performance of different models. In terms of cumulative BLEU-4 score, BERT-Deformer comes out on top with a score of 0.419 compared to the baseline Transformer at 0.358 and the previous fusion model at 0.398. This overall improvement is due, to a large extent, to the deformable attention mechanism's ability to accurately translate critical literary terms. In terms of BERTScore, which reflects depth of meaning, the proposed model achieves a score of 0.886, which indicates an improvement of 0.031 over the benchmark Transformer scores and further highlights how the BERT-Deformer includes a deep semantic understanding component and visual context to improve translation accuracy to the original content. The METEOR scores also reflect that the proposed model performs better than all other comparative models at 0.431, affirming the proposed model's effectiveness in selecting synonyms and maintaining fluidity. Overall, the BERT-Deformer has the most coverage area displayed in the radar chart across each metric. This balanced and strong performance reinforces the comprehensive advantages of the proposed fusion architecture, as it works well to examine literary texts in a multimodal context and provides a strong platform for further qualitative analysis.

B. EFFICIENCY AND PERFORMANCE OPTIMIZATION FOR LONG SEQUENCE TRANSLATION

An essential feature of literary artifacts entails the intricate long sentence structure, as well as cross-paragraph context-dependence, which undoubtedly poses a vital challenge regarding the long sequence processing capacity of machine translation models. To evaluate the performance and resilience of the architecture for long sequence translation, this experiment segregates the test set into five increasing intervals based upon the number of lexemes in the original text: 1-50, 51-100, 101-150, 151-200, and >200. The comparisons of the BLEU score trends of each model across different sequence lengths are observed, and the model's capability to model long-distance context dependencies is evaluated via modeling performance degredation. Detailed results are shown in Figure 4.

Figure 4 presents the changes in translation accuracy as a function of increasing sequence length. The horizontal axis represents the sequence length intervals divided by the number of lemmas, and the vertical axis represents the corresponding BLEU score. It can be seen that the baseline Transformer model shows dramatic degradation in performance after a sequence length of 100 lexical units, as performance drops from 0.385 to score of 0.258, a reduction of 0.127 points. Performance degradation is attributed to the implications of computational efficiency and information deterioration presented by standard attention mechanisms when handling longer sequences. The text model that integrates the BERT terminologists shows some improvements holding data losses to 0.062 points, indicating the separation of deep semantic representation plays a role in holding at least the basic semantic retention. The proposed BERT-Deformer model shows the most desired stability under its model, with the highest performance sustained across the sequence length intervals. Degradation from short to ultra-long is noted at only 0.023 points. The deformable attention in this study effectively manages the inherent challenges of long sequence modeling through dynamically focusing on key information nodes, confirming the architecture's excellent resilience to managing long-form literary texts.

C. CASES ON DEEP SEMANTIC UNDERSTANDING AND TRANSLATION OPTIMIZATION

The greatest difficulty in literary translation is fundamentally in obtaining a deep semantic understanding of the culturally loaded words and re-expressing that meaning on the surface level. Culturally loaded words contain the complete and possibly unique, cultural connotation and aesthetic concept in question, and an exact referent may be impossible to find in the target language. In order to examine the model's capacity for deep semantic understating and cultural transfer, this section uses three representative Japanese culturally loaded words to conduct case studies. By examining comparative differences in translated output from each model,

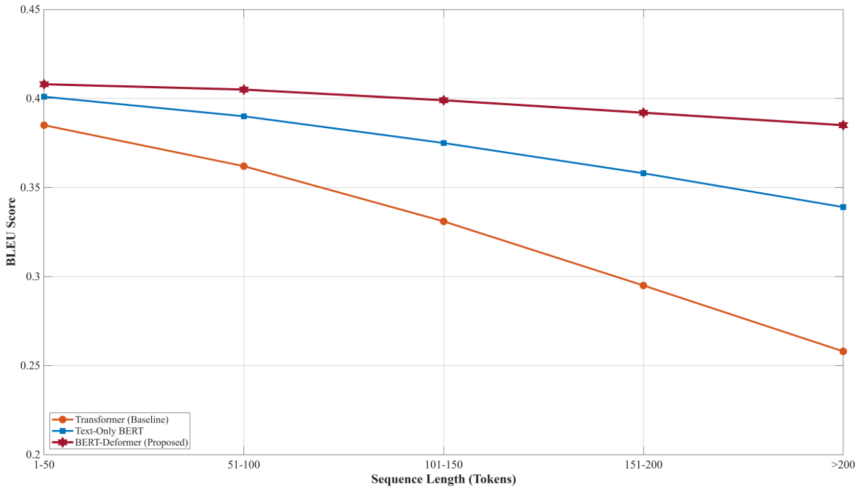

FIGURE 4. Comparison of translation performance for different sequence lengths

and analyzing the results qualitatively, the models should show their strengths and weaknesses in terms of capturing cultural information and artistic expressiveness. Results of the specific translations are shown in Table 4.

Table 4 presents a comparison of translation examples for three typical culturally loaded terms. For the philosophical concept term "もののあはれ", the Transformer baseline model produces a literal translation of "the sadness of things", capturing only the surface emotion while losing its aesthetic essence. The proposed model translates it as "a poignant awareness of impermanence", accurately capturing the core meaning of the fusion of the philosophy of impermanence and sentimental aesthetics. This is attributed to BERT's deep semantic encoding's ability to analyze complex concepts. When processing the aesthetic concept "わびさび", the text-only BERT model outputs a generalized translation of "simple and quiet beauty", failing to reflect the profound connotations of accumulated time. The proposed model, leveraging multimodal context, generates "austere and weathered beauty", where the word "weathered" precisely conveys the temporal dimension of "quiet", demonstrating that visual information helps in understanding abstract aesthetic qualities. When translating the natural phenomenon "むらさめ", the baseline model mistranslates it as "village rain", while the model in this paper, combined with the text and image context, outputs "a sudden, intense rain squall". The use of the technical term "squall" demonstrates that the deformable attention mechanism effectively connects the textual description with visual meteorological features. These examples consistently show that the fusion architecture proposed in this paper significantly improves the translation accuracy and cultural fidelity of culturally loaded words through deep semantic understanding and cross-modal information complementarity.

D. CROSS-MODAL ATTENTION ALLOCATION OF DIFFERENT WORD CATEGORIES

To verify the working mechanism and fusion effectiveness of the cross-modal attention mechanism in the architecture, this section analyzes the dynamic allocation of attention between text and image modalities when processing different types of words. Experiments are conducted by collecting and classifying the decoder attention weights of a large number of translation instances in the test set, calculating the mean cross-modal attention weights corresponding to different part-of-speech categories. The image attention weight for a word represents the proportion of total cross-modal attention allocated to the visual modality when generating that word, ranging from 0 (fully textual) to 1 (fully visual). The system reveals the collaborative mechanism between visual and textual information in the translation process, and the statistical results are shown in Figure 5.

Figure 5 illustrates the cross-modal attention distribution across different parts of speech. The horizontal axis displays the four parts of speech: nouns, culturally loaded nouns, verbs, and adjectives. The vertical axis displays the mean attention weight. Culturally-loaded nouns have the most amount of image attention, with a value of 0.68, while their text attention has a lower value of: 0.32. This means that words in the culturally loaded noun category have the highest degree of visual salience, and the deformable attention mechanism can more easily detect their visually representative attributes in images and further promote image information. Verbs, for example, have noticeably lower image attention (0.35) than text attention (0.65). This is because there are no direct visual cues that can represent the action concepts that verbs express in a static image. For nouns and adjectives, the image attention weight is 0.55 and 0.45, respectively, signifying a moderate reliance of entity concepts and attribute descriptions on multimodal information. The substantial differences in attention distribution across the various parts of speech confirm that the model can adaptively vary its modal fusion strategy depending on

TABLE 4. Qualitative comparison of translations of culturally loaded terms

Japanese original text (culturally loaded words)	Reference translation	Transformer	Text-Only BERT	BERT-Deformer
もののあはれ	the pathos of things	the sadness of things	mono no aware	a poignant awareness of impermanence
わびさび	wabi-sabi	lonely and quiet	simple and quiet beauty	austere and weathered beauty
むらさめ	a sudden rain shower	village rain	a passing rain	a sudden, intense rain squall

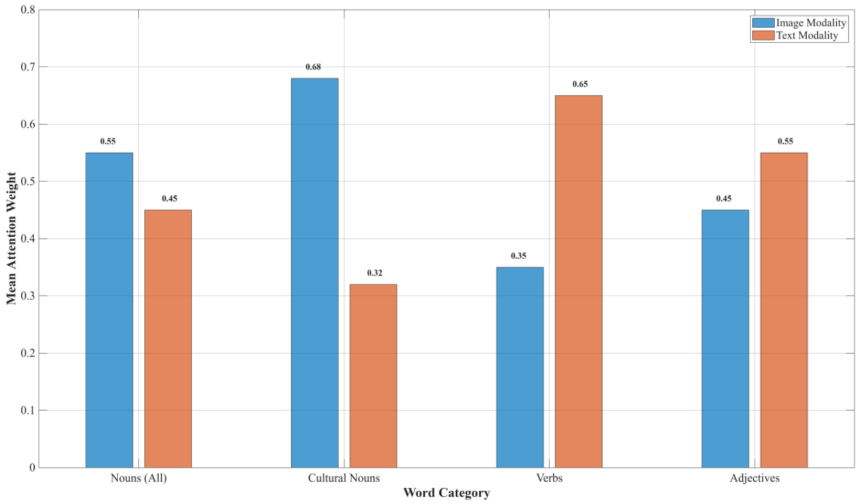

FIGURE 5. Cross-modal attention distribution across word classes

the semantic and syntactic properties of the words. The data-driven attention distribution mechanism provides an important condition for accurate multimodal translating of Japanese literature.

V. CONCLUSION

This paper develops a BERT-Deformer model for multimodal translation of narratives in Japanese literature by leveraging the deep semantic understanding of BERT and the efficient cross-modal fusion ability of a deformable Transformer. The model achieves a BLEU-4 score of 0.419, a METEOR score of 0.431, and a BERTScore of 0.886 on a Japanese literature image-text dataset, with the BLEU-4 score representing an improvement of 0.061 points over the benchmark Transformer model. The BLEU-4 score drops by a mere 0.023 points for long sequences with greater than 200 words, demonstrating strong robustness. The translated culturally loaded nouns achieve an average image attention weight of 0.68, confirming the effectiveness of cross-modal fusion. The experimental results suggest the proposed architecture can successfully parse the complex semantics often found in Japanese literary texts, effectively deploy visual context to enhance semantic disambiguation, and significantly increase the stability during long sequence translations. This work offers an effective technical pathway for handling literary translation tasks rich in culturally loaded words and complex rhetoric, a challenge that mirrors the difficulty in accurately translating historical terminologies and specialized techniques. Consequently, the methodology has clear application value in promoting intercultural communication and digital humanities research, extending its utility to the industry by ensuring the precise cross-cultural transfer of technical and artisanal knowledge crucial for preservation and global trade.

AUTHOR CONTRIBUTIONS

Li Shan designed, collected and analyzed the data, and drafted the manuscript. Li Shan conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Li Shan participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, General Project of Social Science Fund of Liaoning Province in 2023 "Practical Research on the Four in One Linkage Model of Grassroots Party Building in Colleges and Universities to Promote High Quality Talent Cultivation" (Project No. L23BDJ014)

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Naveen and P. Trojovský, "Overview and challenges of machine translation for contextually appropriate translations," Iscience, vol. 27, no. 10, pp. 110878, 2024, doi: 10.1016/j.isci.2024.110878.

[2] L. Gu, "Translation of Japanese Literature Language and Natural Language Environment Understanding Based on Artificial Neural Network," Journal of Environmental and Public Health, vol. 2022, no. 1, pp. 2015763, 2022, doi: 10.1155/2022/2015763.

[3] P. S. Plyth and C. P. Craham, "Translation Affects Literary and Cultural Systems: How to Observe the Features of Translation? Applied Translation," 2023; 14(1):29-37.

[4] G. Corpas Pastor and L. Noriega-Santiáñez, "Human versus neural machine translation creativity: A study on manipulated MWEs in literature," Information, vol. 15, no. 9, pp. 530, 2024, doi: 10.3390/info15090530.

[5] S. M. Abdelhalim, A. A. Alsahil, and Z. A. Alsuhaibani, "Artificial intelligence tools and literary translation: a comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives," Cogent Arts & Humanities, vol. 12, no. 1, pp. 2508031, 2025, doi: 10.1080/23311983.2025.2508031.

[6] C. Herold and H. Ney, "Improving long context document-level machine translation," preprint arXiv, vol. 23, no. 06, pp. 05183, 2023, doi: 10.18653/v1/2023.codi-1.15.

[7] L. Jin, J. He, J. May, and X. Ma, "Challenges in context-aware neural machine translation," preprint arXiv, vol. 23, no. 05, pp. 13751, 2023, doi: 10.18653/v1/2023.emnlp-main.943.

[8] T. Yao, S. Peng, L. Wang, Y. Li, and Y. Sun, "Cross-modality interaction reasoning for enhancing vision-language pre-training in image-text retrieval," Applied Intelligence, vol. 54, no. 23, pp. 12230-12245, 2024, doi: 10.1007/s10489-024-05823-1.

[9] L. Xiao, X. Wu, S. Yang, J. Xu, J. Zhou, and L. He, "Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis," Information Processing & Management, vol. 60, no. 6, pp. 103508, 2023, doi: 10.1016/j.ipm.2023.103508.

[10] J. Wang, "Research on cultural translation based on neural network," Mathematical Problems in Engineering, vol. 2022, no. 1, pp. 6330814, 2022, doi: 10.1155/2022/6330814.

[11] S. Araghi and A. Palangkaraya, "The link between translation difficulty and the quality of machine translation: a literature review and empirical investigation," Language Resources and Evaluation, vol. 58, no. 4, pp. 1093-1114, 2024, doi: 10.1007/s10579-024-09735-x.

[12] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," Information, vol. 15, no. 9, pp. 517, 2024, doi: 10.3390/info15090517.

[13] X. Xu, "Research on neural network machine translation model based on entity tagging improvement," Mathematical Problems in Engineering, vol. 2022, no. 1, pp. 8407437, 2022, doi: 10.1155/2022/8407437.

[14] A. Feng, I. Li, Y. Jiang, and R. Ying, "Diffuser: Efficient Transformers with Multi-hop Attention Diffusion for Long Sequences," arXiv e-prints, vol. 22, no. 10, pp. 11794, 2022.

[15] D. François, M. Saillot, J. Klein, T. F. Bissyandé, and A. Skupin, "Drop-in efficient self-attention approximation method," Machine Learning, vol. 114, no. 6, pp. 139, 2025, doi: 10.1007/s10994-025-06768-3.

[16] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: A comparative analysis," Sensors, vol. 23, no. 5, pp. 2381, 2023, doi: 10.3390/s23052381.

[17] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," ACM computing surveys, vol. 56, no. 9, pp. 1-36, 2024, doi: 10.1145/3649447.

[18] H. Yamauchi, T. Kajiwara, M. Katsurai, I. Ohmukai, and T. Ninomiya, "A Japanese Masked Language Model for Academic Domain," Proceedings of the Third Workshop on Scholarly Document Processing, pp. 152-157, 2022.

[19] E. Odle, Y. J. Hsueh, and P. C. Lin, "Semantic positioning model incorporating bert/roberta and fuzzy theory achieves more nuanced japanese adverb clustering," Electronics, vol. 12, no. 19, pp. 4185, 2023, doi: 10.3390/electronics12194185.

[20] A. R. Sajun, I. Zualkernan, and D. Sankalpa, "A historical survey of advances in transformer architectures," Applied Sciences, vol. 14, no. 10, pp. 4316, 2024, doi: 10.3390/app14104316.

[21] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," Ai, vol. 4, no. 1, pp. 54-110, 2023, doi: 10.3390/ai4010004.

[22] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," Visual computing for industry, biomedicine, and art, vol. 6, no. 1, pp. 14, 2023, doi: 10.1186/s42492-023-00140-9.

[23] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," Technologies, vol. 13, no. 1, pp. 32, 2025, doi: 10.3390/technologies13010032.

[24] Y. Liu, D. Liu, and S. Zhu, "Bilingual-Visual Consistency for Multimodal Neural Machine Translation," Mathematics, vol. 12, no. 15, pp. 2361, 2024, doi: 10.3390/math12152361.

[25] E. Tian, Z. Zhu, F. Liu, Z. Li, R. Gu, and S. Zhao, "Multimodal Machine Translation Based on Enhanced Knowledge Distillation and Feature Fusion," Electronics, vol. 13, no. 15, pp. 3084, 2024, doi: 10.3390/electronics13153084.

[26] Y. Gan, Y. Fu, D. Wang, and Y. Li, "A novel approach to attention mechanism using kernel functions: Kerformer," Frontiers in Neurorobotics, vol. 17, pp. 1214203, 2023, doi: 10.3389/fnbot.2023.1214203.

[27] X. Song, Y. Tian, H. Liu, L. Wang, and J. Niu, "PPLA-Transformer: An Efficient Transformer for Defect Detection with Linear Attention Based on Pyramid Pooling," Sensors, vol. 25, no. 3, pp. 828, 2025, doi: 10.3390/s25030828.

[28] Y. Lu, J. Zeng, J. Zhang, S. Wu, and M. Li, "Learning confidence for transformer-based neural machine translation," arXiv preprint arXiv, vol. 22, no. 03, pp. 11413, 2022, doi: 10.18653/v1/2022.acl-long.167.

[29] T. Pearce, A. Brintrup, and J. Zhu, "Understanding softmax confidence and uncertainty," arXiv preprint arXiv, vol. 21, no. 06, pp. 04972, 2021.

[30] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, et al., "A survey on evaluation metrics for machine translation," Mathematics, vol. 11, no. 4, pp. 1006, 2023, doi: 10.3390/math11041006.

[31] L. Yating, M. Afzaal, X. Shanshan, and D. A. S. El-Dakhs, "TQFLL: a novel unified analytics framework for translation quality framework for large language model and human translation of allusions in multilingual corpora," Automatika, vol. 66, no. 1, pp. 91-102, 2025, doi: 10.1080/00051144.2024.2447652.

[32] Y. Cui and M. Liang, "Automated scoring of translations with BERT models: Chinese and English language case study," Applied Sciences, vol. 14, no. 5, pp. 1925, 2024, doi: 10.3390/app14051925.

[33] Z. Xiao, X. Ning, and M. J. M. Duritan, "BERT-SVM: A hybrid BERT and SVM method for semantic similarity matching evaluation of paired short texts in English teaching," Alexandria Engineering Journal, vol. 126, pp. 231-246, 2025, doi: 10.1016/j.aej.2025.04.061.

[34] X. Zhang, W. Li, X. Wang, L. Wang, F. Zheng, L. Wang, et al., "A fusion encoder with multi-task guidance for cross-modal text-image retrieval in remote sensing," Remote Sensing, vol. 15, no. 18, pp. 4637, 2023, doi: 10.3390/rs15184637.

[35] U. R. Pol, P. S. Vadar, and T. T. Moharekar, "Hugging face: revolutionizing AI and NLP," International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 8, pp. 1121-1124, 2024, doi: 10.22214/ijraset.2024.64023.