Quantification of CITTA Chinese Translation Interpretation Differentiation based on BEiT-RoBERTa Model

XIAOMING CHEN
Faculty of Humanities and Social Sciences,
Khon Kaen University, Khon Kaen 40002,
THAILAND

also with
[2]Kelin Academy, Guangzhou 510000,
Guangdong 510000,
CHINA

Abstract: - This study analyzes varied Chinese renderings of the Buddhist core concept CITTA via quantitative modeling to expose cross-cultural translation discrepancies. It constructs a BEiT-RoBERTa fusion model combining BEiT visual feature extraction and RoBERTa linguistic encoding. Based on a self-built CITTA Chinese translation text dataset containing ancient and modern translations, experiments were conducted from four dimensions: vocabulary recognition, model comparison, semantic coherence, and multi-context adaptation. Performance tests were conducted using indicators such as character error rate (CER), syllable recognition accuracy (SRA), line recognition accuracy (LRA), and manual evaluation. Results show the syllable-based model performs best: CER=0.167, SRA=0.850, LRA=0.791, surpassing CRNN and SRN. The overall accuracy of sentence semantic coherence judgment is 92.68%, and the translation effect in different contexts is between professional translators and non professional enthusiasts, with stable accuracy and fluency. Research has confirmed that the BEiT-RoBERTa model quantifies CITTA translation variations but poorly captures philosophical depth and cultural subtleties.

Key-Words: - CITTA, BEiT-RoBERTa model, Natural language processing, Cross cultural translation, Buddhist philosophy, Philosophical Text Translation.

Received: March 1, 2026. Revised: May 3, 2026. Accepted: June 5, 2026. Published: August 6, 2026.

1 Introduction

The philosophy of spiritual transmission (CITTA) is a core component of the Buddhist philosophical system, and its core concept "mind" (in Pali and Sanskrit: चात्त, Citta) is transliterated qualitatively, not a literal translation referring to the physical heart organ, but rather a highly abstract representation of human thought, mind, intellect, and discernment ability. In the early Buddhist context, "heart" was often used interchangeably with "consciousness" and "meaning", although semantically similar, there were subtle differences, and its connotation continued to enrich with the development of Buddhism.

However, CITTA-related classics and ideas face prominent issues of differential interpretation in Chinese translation during cross-cultural dissemination. Due to significant differences in the language systems, cultural backgrounds, and religious cognition between Chinese and Sanskrit, Chinese translated texts from different periods and translators exhibit obvious differences in vocabulary selection, semantic communication, and philosophical connotation interpretation [1], [2] - including differences in literal expression, as well as misunderstandings of core philosophical concepts, posing challenges to the accurate dissemination and in-depth research of CITTA ideas, [3]. At present, academic research on such differentiation primarily relies on qualitative analysis, lacking objective and systematic quantitative tools, making it difficult to accurately capture the subtle changes and big differences in semantic transmission during the Chinese translation process, and unable to provide a scientific basis for evaluating and optimizing translation quality.

To address this issue, this study introduces natural language processing technology and constructs a BEiT-RoBERTa fusion model that integrates BEiT visual feature extraction and RoBERTa language modeling. Quantitative research is conducted on the differentiation of CITTA

Chinese translation interpretation. This model combines the accuracy of visual feature extraction with the semantic understanding ability of language modeling and can objectively quantify the translation of ancient and modern CITTA texts from dimensions such as vocabulary recognition, semantic coherence, and multi-context adaptation. In recent years, end-to-end OCR and translation fusion models based on the Transformer architecture have made significant progress in multilingual text recognition and translation tasks. However, most existing models are aimed at general scenario texts or standardized documents, with training data mainly consisting of large-scale general parallel corpora, [4]. The optimization goal focuses on character-level and word-level translation and recognition accuracy, and there is no specialized modeling and quantitative evaluation for the interpretation differences of core concepts in religious and philosophical classics. Compared to existing models, the innovation of this study mainly lies in two aspects: firstly, in terms of task objectives, it does not focus solely on maximizing translation fidelity or recognition accuracy, but instead conducts a systematic quantitative analysis for the first time on the differences in interpretation of the core concept of Buddhist philosophy CITTA in Chinese translation; Secondly, in terms of model architecture, BEiT-RoBERTa does not simply stack visual and language modules, but designs syllable level recognition units and cross attention semantic alignment mechanisms for religious texts with low resource and high ambiguity characteristics. While maintaining end-to-end trainability, it enhances the perception of concept drift.

In summary, the aim of this study is to validate the applicability of the BEiT-RoBERTa model in the quantitative analysis of religious philosophy texts, providing a scientific evaluation tool and theoretical support for the translation of Buddhist scriptures, and thereby enhancing the accuracy and effectiveness of cross-cultural dissemination of Buddhist philosophy.

2 Theoretical Basis and Experimental Design

2.1 The Core Concept of the Philosophy of the Transmission of the Soul

CITTA, as a key component of the Buddhist philosophical system, contains profound and unique ideological connotations. The core concept of "mind" (चित्त, Citta) is the cornerstone of the entire philosophical system. This concept is not simply equivalent to the modern sense of the heart organ, but covers multiple levels such as thought, consciousness, emotion, and cognition, and is a highly abstract and generalized representation of the human spiritual world.

CITTA occupies an extremely important position in the grand field of religious philosophy. It provides theoretical guidance for Buddhist practitioners to deeply explore their inner world and gain insight into the essence of life. Through the study of the essence and activity patterns of the "heart", Buddhists attempt to achieve spiritual liberation and transcendence. With the widespread spread of Buddhism worldwide, CITTA is also facing challenges in cross-cultural communication and translation. People from different linguistic and cultural backgrounds inevitably have differences in understanding and interpreting CITTA. These differences are not only due to differences in language expression but also closely related to the philosophical concepts, value orientations, and ways of thinking contained in various cultures.

How to accurately convey the core concepts and essence of CITTA in the translation process has become an urgent problem to be solved. Due to the differences in the connotation and extension of the concept of "heart" in different cultures, a simple literal translation often cannot accurately express its rich meaning. For example, in Chinese, "heart" can represent both the physical heart organ and spiritual aspects such as thoughts, emotions, and willpower; In English, the word "heart" corresponds to "heart", although having a similar meaning to some extent, has significant differences in cultural context and semantic emphasis from the Chinese word 'heart'. Therefore, when translating CITTA-related classics, it is necessary to fully consider these cultural differences and adopt appropriate translation strategies and methods to ensure that the translation can accurately convey the philosophical ideas and cultural connotations of the original text.

2.2 RoBERTa Model Principle

The RoBERTa model is a pre trained language model based on the Transformer architecture, which has demonstrated excellent performance and powerful influence in the field of natural language processing, [5], [6]. The model architecture is mainly composed of multiple stacked Transformer blocks, which are the core components of the RoBERTa model, responsible for feature extraction and semantic understanding of input text.

The self attention mechanism is a key component of the RoBERTa model, which breaks the limitations of traditional recurrent neural networks (RNN) and

convolutional neural networks (CNN) in processing sequence data. It can effectively capture the dependency relationships between different positions in the input sequence, thereby better understanding the contextual information of the text. In general, the self attention mechanism determines the importance of each position in the current context by calculating the attention weight between each position in the input sequence and other positions. In this way, the model can obtain relevant information from the entire input sequence based on the query vector of the current location, thereby achieving effective modeling of long-distance dependencies. Therefore, this study proposes an end-to-end text recognition method, BEiT-RoBERTa, that integrates pre-trained models BEiT [7] and RoBERTa [8], which can recognize longer texts with more complex glyph patterns and higher recognition difficulty.

2.3 BEiT-RoBERTa Model

2.3.1 Overall Framework of BEiT-RoBERTa

The BEiT-RoBERTa model architecture used in this study is based on the Transformer encoder-decoder structure. The overall framework of the model is illustrated in Figure 1. The encoder part of the model adopts the BEiT algorithm to extract visual features of images, while its decoder part performs language modeling tasks based on the RoBERTa algorithm, [9].

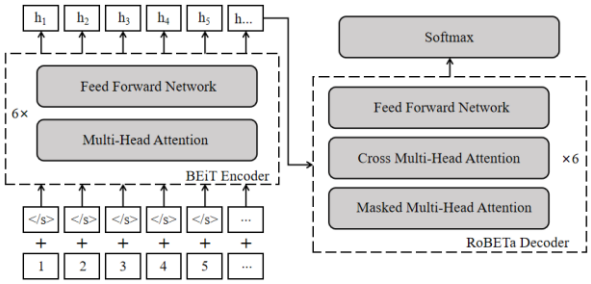

Fig. 1: BEiT-RoBERTa model architecture
Source: created by the authors

In the process of model training, the input image is first cut into several 16×16-pixel image blocks, and then special symbols and corresponding position codes are added to form the input sequence of the model. Secondly, utilize the pre-trained BEiT encoder to output serialized hidden vectors. Furthermore, the labels of the text image are divided into syllables as recognition units, and the resulting divisions are simultaneously input to the RoBERTa decoder along with the encoder output. Finally, the Softmax layer is used to predict and output at each time step, and all predicted results are concatenated in sequence to generate the translation results of the text line images.

2.3.2 Encoder BEiT

In the process of feature extraction from the original image using a pre-trained BEiT model, the height of the image $f \in R^{H\times W\times 3}$ is first uniformly cropped to 64; Then, the fixed-size image is uniformly segmented into 16×16 blocks, and each segmented block is flattened and mapped to a D-dimensional space to complete the block embedding operation of the image. Subsequently, the image feature sequence is overlaid with positional encoding to generate the final input of the encoder.

The structure of the BEiT encoder is shown in Figure 2, consisting of six identical layers stacked together, each employing unique parameter weights. The layer structure includes two core submodules: multi-head attention and Feed forward networks (FFN). After each submodule, residual structure and layer normalization are immediately applied.

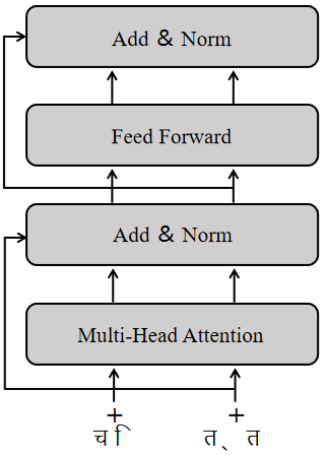

Fig. 2: BEiT encoder
Source: created by the authors

When implementing the multi head attention mechanism, this study adopted scaled dot product attention to structurally process sequence data, with the specific expression as Equation (1):

$$Attention(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (1)$$

Among them, Q is the query matrix, K is the key matrix, V is the value matrix, and dK is the scaling factor.

Multi-head attention is essentially the process of multiplying attention by scaling points H times and integrating the resulting outputs. The formula for the multi head attention mechanism is Equations (2) and (3):

$$MultiHead(Q,K,V)=Concat(head_1,...,head_h)W^o \quad (2)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (3)$$

Among them, W^o, W^Q, W^K, and W^V represent corresponding weight vectors, all of which belong to trainable parameters; h represents the number of heads.

As a fully connected network, a feedforward neural network consists of two fully connected layers, where the first layer uses ReLU as the activation function and the second layer does not use an activation function. The expression is Equation (4):

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2 \quad (4)$$

Among them, x is the output after multi-head attention, and W_1, W_2, b_1, and b_2 are all trainable parameters.

Given that multi-layer attention mechanisms may result in loss of low-level feature information, this study introduces residual connections and layer normalization after each sublayer to ensure that the encoder output vector still retains the original feature information. The calculation method for the output of the sublayer is Equation (5):

$$LayerNorm(x+Sublayer(x)) \quad (5)$$

Among them, x is the output of the previous sublayer, representing the original input data only in the first sublayer of the first layer.

2.3.3 Decoder RoBERTa

During the decoding phase, RoBERTa is used, and its pre-training results were utilized to initialize the decoder. The RoBERTa model is an improvement upon the BERT model, and its key adjustments include: firstly, removing the next sentence prediction target; The second is to dynamically change the masking mode of the masked language model, [10].

Due to the lack of a cross-attention mechanism in the RoBERTa structure, direct interaction between the encoder and decoder cannot be formed. To address this issue, this study introduces a cross-attention layer between masked multi-head attention and feedforward neural networks, thereby achieving attention allocation between the encoder output and decoder input, and linking them with the decoder output. Directly importing the pre-trained model into the decoder during loading can result in structural mismatches. Therefore, this study adopted a strategy of initializing the decoder using a pre-trained RoBERTa model and randomly initializing the cross-attention layer. The specific structure of the decoder is shown in Figure 3.

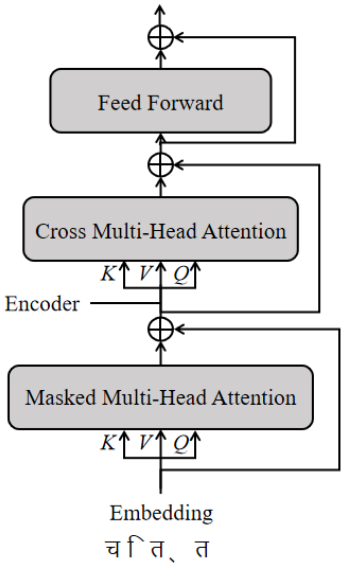

Fig. 3: RoBERTa decoder
Source: created by the authors

The input content of the decoder mainly consists of two parts: one is the result of word embedding in the target text, and the other is the output of the encoder. These two pieces of information are respectively fed into the multi-head attention layer and cross-attention layer of the decoder. The first attention layer is a masked multi-head attention layer, and the attention score will be calculated by multiplying the scaled points by the attention. Unlike the attention mechanism in the encoder, the attention mechanism in the decoder introduces the Mask mechanism. In the second cross-attention sublayer, the scaling point by the attention mechanism is still used to calculate scores. This process utilizes the output K and V matrices of the encoder, as well as the Q matrix generated by the previous mask attention layer, to achieve simultaneous attention to the current decoding input and the previous encoded output, thereby integrating local and global attention. Finally, the feedforward neural network layer takes the outputs of the masked multi-head attention layer and the cross self-attention layer as inputs and performs nonlinear transformation and mapping through a multi-layer perceptron structure.

The final output of the decoder goes through the Softmax layer, which converts the output of the neural network into probability values and outputs the probability distributions of each category, thereby completing the text translation task.

2.4 Experimental Dataset

The experiment was conducted on a self-constructed CITTA Chinese translation text dataset, which mainly covers important translation achievements from ancient and modern times. The ancient translation texts selected relevant translations by famous translators such as Xuanzang and Kumarajiva, while the modern translation texts collected new translations by multiple contemporary scholars. The performance of the BEiT-RoBERTa model is measured using evaluation metrics such as character error rate (CER), syllable recognition accuracy (SRA), and line recognition accuracy (LRA).

In the dataset, first, randomly select 10 syllables as units; Then, generate a text line image of CITTA; Finally, 2 million text line images were obtained to form the CITTA Chinese translation text dataset. During the experiment, 5% of the CITTA dataset was randomly selected as the validation set. The height of each text image is 64 pixels, and the width is not fixed.

2.5 Evaluation Indicators

Evaluate the recognition performance using CER, SRA, and LRA; Cohen's Kappa coefficient is used to evaluate the consistency of judgments.

CER compares the predicted character sequence output with the correct reference character sequence using the Levenshtein distance, which calculates the number of character insertions i, substitutions s, and deletions d, and normalizes them based on the total number of characters n in the text. The specific calculation method is Equation (6):

$$CER = \left[\left(i + s + d\right)/ n\right] \times 100 \tag{6}$$

SRA reflects the accuracy of syllable recognition in the prediction results, covering situations of incorrect recognition, missed recognition, and over-recognition, accounting for the proportion of all CITTA syllables. Its calculation method is Equation (7):

.

$$SRA = \sum_{i=1}^{N} \frac{\left[len\left(P_i\right) - edit\left(P_i, L_i\right)\right]}{len\left(P_i\right)} / N \tag{7}$$

Among them, N is the number of predicted text line images, P_i is the CITTA syllable in the predicted result, L_i is the syllable in the label, and edit is the edit distance between the predicted result and the real label.

LRA refers to the ratio of correctly predicted text lines to the total text lines, calculated as in Equation (8):

$$LRA = \frac{\sum_{j=1}^{N}\left(P_j = L_j\right)}{\sum_{j=1}^{N} L_j} \tag{8}$$

Among them, P_j is the predicted result of the text line, and L_j is the corresponding label of the text line. If there is a difference between the predicted result P_j and the label L_j, it is judged as a prediction error.

Cohen's Kappa coefficient is a statistical indicator used to measure the degree of consistency between two evaluators. The specific calculation method is Equation (9):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

Among them, p_o is the observation consistency rate, and p_e is the accidental consistency rate.

2.6 Hardware Configuration and Model Parameters

The experiment used the deep learning framework PyTorch to build the model BEiT-RoBERTa, and trained the model using the RTX 2080 Ti GPU.

The parameter configuration of the BEiT-RoBERTa model is as follows: using a vision encoder decoder, the encoder part adopts a 6-layer Transformer structure, with a hidden layer size of 384, 6 attention heads, and a dimension of 1536 for the feedforward neural network; The decoder part also has 6 layers, 8 attention heads, 384 hidden layers for cross attention, and 1024 dimensions for the feedforward neural network. The total number of parameters in the model is about 147M, of which the encoder accounts for about 62M and the decoder accounts for about 85M.

During the training process, the loss function rapidly decreases in the first cycle, dropping from an initial 2.34 to 0.67; After the third cycle, the decline tends to flatten out, and after the fifth cycle, the CER on the validation set begins to stabilize around 0.17, with SRA and LRA converging to around 0.85 and 0.79, respectively. No significant overfitting was observed during the training process, and the performance gap between the validation set and the training set remained within 3%. Using Adam optimizer, with an initial learning rate of 0.00005.

3 Experimental Results

3.1 Recognition of Chinese Vocabulary Translation

In this study, we employed end-to-end BEiT-RoBERTa technology to validate the CITTA Chinese translation text dataset. The evaluation was conducted through CER, SRA, and LRA, and the experimental results are shown in Table 1. From the analysis of the data in the table, it can be seen that there is a decreasing trend from letter to syllable; On the other hand, SRA and LRA exhibit an increasing state. Based on this inference, the recognition method based on syllables exhibits the best performance in CITTA translation recognition tasks. This phenomenon may be attributed to the syllable-level text segmentation strategy, which effectively avoids the problem of cohesion between characters.

Table 1. Recognition results of different lexical types

Word type	CER	SRA	LRA
Letter	0.180±0.005	0.825±0.008	0.492±0.015
Font	0.175±0.003	0.833±0.006	0.691±0.012
Syllable	0.167±0.004	0.850±0.005	0.791±0.012

Source: created by the authors

3.2 Comparison with Mainstream Methods

To verify the superiority of the constructed model BEiT-RoBERTa, it was compared with the methods proposed in previous studies. This paper selected the BEiT-RoBERTa model and current popular text recognition algorithms to conduct comparative experiments on the independently constructed CITTA Chinese translation text dataset. In the field of text recognition, CRNN and SRN are commonly used recognition techniques. In view of this, this study selected CRNN and SRN as control algorithms, and the specific results of the experiment are shown in Table 2.

Table 2. Comparison of different methods

Method	Conceptual vocabulary		
	CER	SRA	LRA
CRNN	0.180±0.007	0.845±0.009	0.768±0.011
SRN	0.183±0.009	0.842±0.010	0.752±0.013
BEiT-RoBERTa	0.167±0.006	0.850±0.007	0.791±0.009

Source: created by the authors

After comparing and analyzing the recognition performance of SRN [11], CRNN, and BEiT-RoBERTa, it can be observed that the model proposed in this study achieves better performance than current recognition models in text recognition, with CER, SRA, and LRA even showing better results than existing methods.

3.3 Evaluation of Sentence Semantic Coherence

To evaluate the understanding ability of the BEiT-RoBERTa model on the semantic coherence of CITTA Chinese-translated sentences. 50 sentences were randomly selected from the CITTA Chinese translation text dataset, and Buddhist researchers manually annotated the semantic coherence of the sentences based on their semantic content and logical relationships, dividing them into three levels: "coherent", "basic coherent", and "incoherent". Then, these sentences are input into the RoBERTa model, which judges the semantic coherence of the sentences by analyzing the semantic associations, syntactic structures, and contextual information between the vocabulary in the sentences. A sentence S is composed of n words, and its semantic coherence is defined as the conditional probability product between adjacent semantic units, i.e. $P_{coh}(S)=\prod_{i=2}^{n} P(w_i|w_{i-1},\theta_{RoBERTa})$, where P(.) is the conditional probability of the model generating the current word given the previous word. The global coherence score is defined as $C(S)=\frac{1}{n-1}\sum_{i=2}^{n}\log P(w_i|w_{i-1},\theta_{RoBERTa})$, and the higher the score, the smaller the semantic jump and the more coherent the sentence. Consider the manually annotated coherence level as a latent variable, and the model's judgment result as an observed variable. Construct a confusion matrix and calculate Cohen's Kappa coefficient based on annotated data of 50 sentences to measure the consistency between the model and manual judgment.

According to the results in Table 3, κ=0.872 indicates a significant consistency between the model's judgment and that of human experts; the BEiT-RoBERTa model reached 92.68% in determining sentence semantic coherence. For coherent sentences, the accuracy of the model's judgment is 94.44%, which can accurately identify the close connections between vocabulary and semantics in the sentence, as well as the complete logic expressed by the sentence. For basic coherent sentences, the accuracy of the model is 93.33%. For incoherent sentences, the accuracy is relatively low at 87.50%. This may be because incoherent sentences often have problems such as semantic jumps and logical contradictions, which increase the difficulty of model understanding and judgment.

Overall, the BEiT-RoBERTa model performs well in evaluating sentence semantic coherence and can effectively assist in analyzing the quality and accuracy of CITTA-translated sentences.

Table 3. Evaluation results of sentence semantic coherence

Coherence level	Manually labeled quantity	Model judgment quantity	Accuracy	Kappa
Coherent	18	17	94.44	0.31
Basic coherence	15	14	93.33	0.28
Incoherent	17	17	87.50	0.28
Overall	50	48	92.68	k=0.872

Source: created by the authors

3.4 Comparison of Translation Effects in Different Contexts

In order to compare the performance of the BEiT-RoBERTa model in CITTA translation in different contexts, three different contexts were set up, namely the Buddhist classic context, the modern academic research context, and the popular culture context. For each context of the text, it is translated separately by professional translators, translation enthusiasts, and the BEiT-RoBERTa model. After the translation is completed, invite Buddhist research experts to evaluate the translation results from three dimensions: accuracy, fluency, and cultural adaptability. The scoring range for each dimension is 1-5 points, with 5 points being the highest score. From the data in Table 4, it can be seen that professional translators perform well in accuracy, fluency, and cultural adaptability in different contexts, with a high overall score. The performance of translation enthusiasts in various dimensions is relatively poor, especially in the context of Buddhist scriptures and modern academic research, and there is a significant gap compared to professional translators and the BEiT-RoBERTa model. The translation performance of the BEiT-RoBERTa model in different contexts is between that of professional translators and non-professional translation enthusiasts. In terms of accuracy, the BEiT-RoBERTa model can better understand the meaning of the original text and provide more accurate translations. In terms of fluency, the translations generated by the model also have a certain degree of fluency, but there is still room for improvement in cultural adaptability compared to professional translators. In the context of Buddhist scriptures, due to the involvement of a large number of professional terms and specific cultural background knowledge, the shortcomings of the BEiT-RoBERTa model in terms of cultural adaptability are more pronounced; In the context of popular culture, the performance of the model is relatively good because the language expression in the context of popular culture is more easy to understand and relies relatively less on cultural background knowledge.

Table 4. Comparison of translation effectiveness in different contexts

Context	Translation subject	Accuracy	Fluency	Cultural adaptability	Comprehensive
Buddhist classic context	Professional translation	4.5	4.2	4.3	4.33
	Translation enthusiasts	3.0	2.8	2.5	2.77
	BEiT-RoBERTa	4.0	3.8	3.5	3.77
Context of modern academic research	Professional translation	4.3	4.0	4.1	4.13
	Translation enthusiasts	2.8	2.5	2.3	2.53
	BEiT-RoBERTa	3.8	3.5	3.2	3.50
Mass cultural context	Professional translation	4.0	4.2	4.0	4.07
	Translation enthusiasts	3.2	3.0	2.8	3.00
	BEiT-RoBERTa	3.5	3.8	3.3	3.53

Source: created by the authors

4 Discussion

Through four experimental results, including vocabulary translation recognition, comparison of mainstream methods, evaluation of sentence semantic coherence, and comparison of translation

effects in different contexts, it can be seen that the BEiT-RoBERTa model constructed in this study has significant value in the quantitative analysis of CITTA Chinese translation interpretation differentiation. Its performance advantages and application limitations can be explored from multiple dimensions. At the level of vocabulary recognition, the BEiT-RoBERTa model performs the best with a syllable-based recognition strategy, achieving the lowest CER and highest SRA, significantly better than letter-level and character-level recognition. This result is attributed to the fusion of BEiT's visual feature extraction ability and RoBERTa's language modeling advantage in the model. Syllable-level segmentation effectively avoids recognition interference caused by character adhesion, while also fitting the language structure features of CITTA text. However, it should be noted that for core terms in Buddhist philosophy that have multiple interpretive dimensions, the model can only achieve accurate recognition at the literal level, making it difficult to directly capture their big philosophical differences. Further interpretation should be combined with specialized knowledge in religious studies.

In the comparison of model performance, BEiT-RoBERTa outperforms traditional CRNN and SRN algorithms in terms of CER, SRA, and LRA on the independently constructed CITTA Chinese translation dataset, verifying the adaptability of the visual language fusion architecture in religious text recognition tasks. This advantage stems from two aspects: firstly, the multi-head attention mechanism and residual connection design of the encoder BEiT effectively preserve the original feature information of the text image; The second is that the decoder RoBERTa establishes a direct interaction between encoding and decoding processes through the introduction of cross-attention layers, improving the accuracy of semantic mapping, [12]. However, the comparison results also show that the performance gap between the model and mainstream methods is not extremely significant, indicating that traditional algorithms still have certain applicability in scenarios with high text clarity and a relatively single context. The core value of BEiT-RoBERTa lies in handling translation recognition tasks in complex character shapes, long texts, and multiple contexts.

In terms of sentence semantic coherence judgment, the overall accuracy of the model is 92.68%, especially for sentences with complete logic and tight semantics, the recognition accuracy exceeds 94%, reflecting its deep capture ability of syntactic structure and contextual correlation. This provides an efficient tool for evaluating the quality of CITTA Chinese-translated texts, which can quickly filter out translated fragments with semantic distortion. However, when dealing with sentences with semantic jumps and logical implications, the accuracy of the model decreased to 87.5%, reflecting its insufficient ability to handle special semantic forms such as metaphorical expressions and Zen-like expressions in Buddhist texts. This type of text often relies on a deep combination of spiritual experience and cultural context, and it is difficult to fully grasp its internal logic solely through language models.

In the comparison of translation effects in different contexts, the comprehensive performance of BEiT-RoBERTa is between professional translators and non-professional enthusiasts, showing stable advantages in accuracy and fluency, but there are obvious shortcomings in the cultural adaptability dimension. In the context of Buddhist scriptures, due to the dense use of professional terminology and profound cultural connotations, the cultural adaptability score of the model is only 3.5 points, significantly lower than the 4.3 points of professional translation; In the context of popular culture, due to the popular language expression and low cultural dependence, the model score has been raised to 3.53 points, significantly narrowing the gap with professional translation. This phenomenon indicates that the integration of religious and cultural knowledge in the current model is still insufficient, making it difficult to accurately convey the deep cultural information contained in the CITTA concept, such as spiritual concepts and sectarian traditions.

5 Conclusion

This study focuses on the quantitative analysis of the differences in interpretation of the CITTA Chinese translation. Taking CITTA, the core concept of Buddhist philosophy, as the research object, a BEiT-RoBERTa visual language fusion model integrating BEiT and RoBERTa was constructed. Through the independently constructed CITTA Chinese translation text dataset, experiments were conducted from four dimensions: vocabulary recognition, model comparison, semantic coherence, and multi-context adaptation. The BEiT-RoBERTa model provides an effective quantitative analysis tool for the study of CITTA Chinese translation differentiation, and its multidimensional evaluation ability helps reveal the differences in vocabulary selection, semantic communication, and contextual adaptation among different translators. But the limitations of the model are also very clear: insufficient capture of deep philosophical connotations, limited ability to handle complex semantic logic, and a need to improve

cultural adaptability. Therefore, in practical applications, it is necessary to combine the quantitative results of the model with manual analysis and professional knowledge of religious studies in order to interpret the differences comprehensively and accurately in the process of CITTA Chinese translation. Future research can expand training data that includes annotations of philosophical connotations, enhancing the model's understanding of the deep meanings of religious concepts; Optimize the parameters of the cross attention layer to enhance the modeling ability for complex semantic logic; Integrating into the Buddhist cultural knowledge base, strengthening the model's ability to adapt to cultural contexts, thereby providing more comprehensive support for cross-cultural translation and research of Buddhist classics.

References:

[1] Smith. Interrogating constructive realism about the self from a Buddhist perspective: Buddhist philosophy and the embodied mind: A constructive engagement, Philosophical Psychology, Vol. 38, No. 4, 2025, pp.1887-1891. doi: 10.1080/09515089.2022.2159798Q.

[2] Barth FG, Giampieri-Deutsch P, Klein HD. Sensory perception, body and mind in Indian Buddhist philosophy. Springer Vienna, Springer, Vienna, 2012, pp.357-368. doi: 10.1007/978-3-211-99751-2_20.

[3] Thakchoe S. Buddhist philosophy of mind: Nāgārjuna's critique of mind-body dualism from his rebirth arguments. Philosophy East and West, Vol. 69, No. 3, 2018, pp.807-827. doi: 10.1353/pew.2019.0064.

[4] Khallouli W, Uddin MS, Sousa-Poza A, Li J, Kovacic S. Leveraging transformer-based ocr model with generative data augmentation for engineering document recognition, Electronics, Vol. 14, No. 1, 2025, 5. doi: 10.3390/electronics14010005.

[5] Jayawardena A. The Meaning of Life: A Perspective from Buddhist Philosophy and the Agganna Sutta. Europe PMC, 2025, 2025010358. doi: 10.20944/preprints202501.0358.v1.

[6] Nguyen VT, Trinh PN. Suffering, authenticity and freedom: comparative perspectives of Buddhist and Heideggerian conceptions of human existence in east-west dialogue. Open Journal of Philosophy, Vol. 15, No. 2, 2025, pp.453-467. doi: 10.4236/ojpp.2025.152026.

[7] Cui X, Song C, Li D, Qu X, Long J, Yang Y, Zhang H. RoBGP: A Chinese nested biomedical named entity recognition model based on RoBERTa and global pointer. Computers, Materials & Continua, Vol. 78, No. 3, 2024, pp.73603-3618. doi: 10.32604/cmc.2024.047321.

[8] Meng Q, Zhang X, Dong Y, Chen Y, Lin D. A combined semantic dependency and lexical embedding RoBERTa model for grid field relational extraction. Applied Sciences-Basel, Vol. 13, No. 19, 2023, p.11074. doi: 10.3390/app131911074.

[9] Ibitoye AOJ, Onifade OFW. Utilizing RoBERTa model for churn prediction through clustered contextual conversation opinion mining. International Journal of Intelligent Systems and Applications, Vol. 15, No. 6, 2023, pp.1-8. doi: 10.5815/ijisa.2023.06.01.

[10] Sobhanam H, Prakash J. Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification. International Journal of Information Technology, Vol. 15, No. 7, 2023, pp.3669-3677. doi: 10.1007/s41870-023-01395-4.

[11] Mujahid M, Kanwal K, Rustam WAI. Arabic ChatGPT tweets classification using RoBERTa and BERT ensemble model. ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 2, No. 8, 2023, pp.1-23. doi: 10.1145/3605889.

[12] Sun J, Chen G. Text sentiment analysis model based on RoBERTa-BiLSTM Attention. In Third International Conference on Advanced Algorithms and Neural Networks (AANN 2023), 2023, Vol. 12791, pp. 343-348. doi: 10.1117/12.3004930.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Xiaoming Chen: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Resources, Software

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.