Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 7 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i07.54979

PAPER

An Innovative Translation Teaching Model Based on Mobile Technology: A Case Study of Translation Major Classrooms

Hongjing Chang, Jing Zhao(✉)

School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

lelezhao0412@gmail.com

ABSTRACT

With the rapid development of information technology, the application of mobile technology in education has become increasingly widespread, particularly in language learning and translation teaching. In translation major classrooms, traditional teaching methods are gradually failing to meet the demands of training translation professionals in the new era. Existing studies have predominantly focused on the use of individual translation tools or basic translation support technologies, lacking systematic and in-depth discussions. Particularly in the context of real-time bilingual translation in teaching scenarios, how mobile technology and advanced machine translation algorithms can be integrated to improve translation efficiency and quality in the classroom remains an unresolved issue. A new innovative translation teaching model based on mobile technology was proposed in this study, with two core aspects being examined. First, to address the need for real-time bilingual translation in translation major classrooms, a hybrid tensor train decomposition (HTTD) method was introduced, which optimizes the flow of information and computational processes in translation tasks through efficient model decomposition and multi-dimensional data fusion. Second, based on HTTD, a lightweight machine translation model was developed, aiming to reduce the computational complexity and resource consumption during the translation process, ensuring the real-time performance and responsiveness of the translation system on mobile devices. This study not only provides a new technical support model for translation teaching but also offers innovative insights for the optimization and application of machine translation systems, holding significant theoretical and practical value.

KEYWORDS

mobile technology, translation teaching, real-time bilingual translation, hybrid tensor train decomposition, lightweight machine translation model

Chang, H.J., Zhao, J. (2025). An Innovative Translation Teaching Model Based on Mobile Technology: A Case Study of Translation Major Classrooms. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(7), pp. 224–238. https://doi.org/10.3991/ijim.v19i07.54979

Article submitted 2024-12-07. Revision uploaded 2025-01-10. Final acceptance 2025-02-13.

1 INTRODUCTION

With the rapid development of mobile technology and the widespread use of smart devices, the field of education is undergoing a profound transformation from traditional classrooms to digital and intelligent teaching models [1–4]. In translation major education, real-time bilingual translation between teachers and students plays a core role in enhancing language application skills and cross-cultural communication competence. However, traditional classrooms are constrained by outdated hardware and teaching tools, making it difficult to achieve efficient, low-latency interactive translation training [5–7]. The introduction of mobile technology offers a new path to address this issue, but its successful implementation requires overcoming the conflict between model computational complexity and resource constraints on mobile devices [8–13]. Constructing a lightweight machine translation model that balances high accuracy with low power consumption has become a key technological challenge in promoting innovation in translation teaching.

This study aims to explore an innovative translation teaching model based on mobile technology, with significance in both theoretical and practical dimensions. On one hand, the localized deployment of lightweight translation models can break through the network dependency and privacy risks of cloud services, providing secure and stable real-time translation support for classroom teaching. On the other hand, technological innovations that optimize teacher-student interaction can effectively stimulate learners' initiative and promote the closed-loop training of translation skills from input to output. Furthermore, this study offers an interdisciplinary paradigm for the deep integration of educational technology and artificial intelligence, providing important reference value for advancing the digital transformation of translation teaching.

Currently, the optimization of machine translation models for mobile devices primarily focuses on model compression and quantization techniques, yet significant limitations remain [14, 15]. While traditional Transformer models demonstrate excellent performance in translation quality, their large parameter size and computational complexity make them ill-suited to the processing power and memory limitations of mobile devices. Existing methods such as low-rank decomposition and knowledge distillation, although capable of reducing model size, often lead to performance degradation due to excessive compression, particularly in the handling of technical terminology and complex sentence structures [16–19]. Furthermore, current research tends to concentrate on general scenarios, with a lack of customized designs that address the real-time interaction and multimodal feedback demands specific to translation teaching contexts, limiting the effectiveness of technology implementation [20–22].

This study addresses the aforementioned issues through two core components: first, a hybrid tensor train decomposition (HTTD) method was proposed for real-time bilingual translation in translation classrooms. By coupling low-rank tensor chains with fully connected layers, the model parameters were compressed while maintaining high-order semantic representation capabilities. Second, based on HTTD, a lightweight Transformer model was developed, which delivered low-latency, high-precision localized translation services, after optimizing the computation process and memory usage in terms of mobile hardware characteristics. This study not only provides a practical technical solution for translation teaching but also contributes theoretical innovation in lightweight model design. The results hold significant practical value for advancing the contextual application and sustainable development of mobile educational technologies.

2 HTTD FOR REAL-TIME BILINGUAL TRANSLATION BETWEEN TEACHERS AND STUDENTS IN TRANSLATION MAJOR CLASSROOMS

This study addresses the application requirements of mobile technology in translation teaching contexts, proposing the construction of a lightweight Transformer model based on HTTD. The core design motivation stems from the dual technological demands of mobile teaching scenarios. Traditional Transformer models exhibit a redundancy in parameters within the self-attention mechanism and fully connected layers, making it difficult to achieve low-latency real-time translation interaction on mobile devices. By introducing HTTD to perform high-order decomposition and reconstruction of weight tensors, the model's storage requirements and computational complexity can be significantly reduced, while preserving the core feature-capturing capability of multi-head attention. Simultaneously, the lightweight Transformer model provides technical support for the development of mobile-embedded translation teaching systems. By integrating the optimized model into mobile devices such as tablets or smartphones, teachers can create hybrid teaching scenarios that blend virtual and physical elements: students can perform real-time bilingual translation training through mobile devices in the classroom, with attention visualization maps and error annotation data generated in real-time by the system providing teachers with precise analytics on student performance.

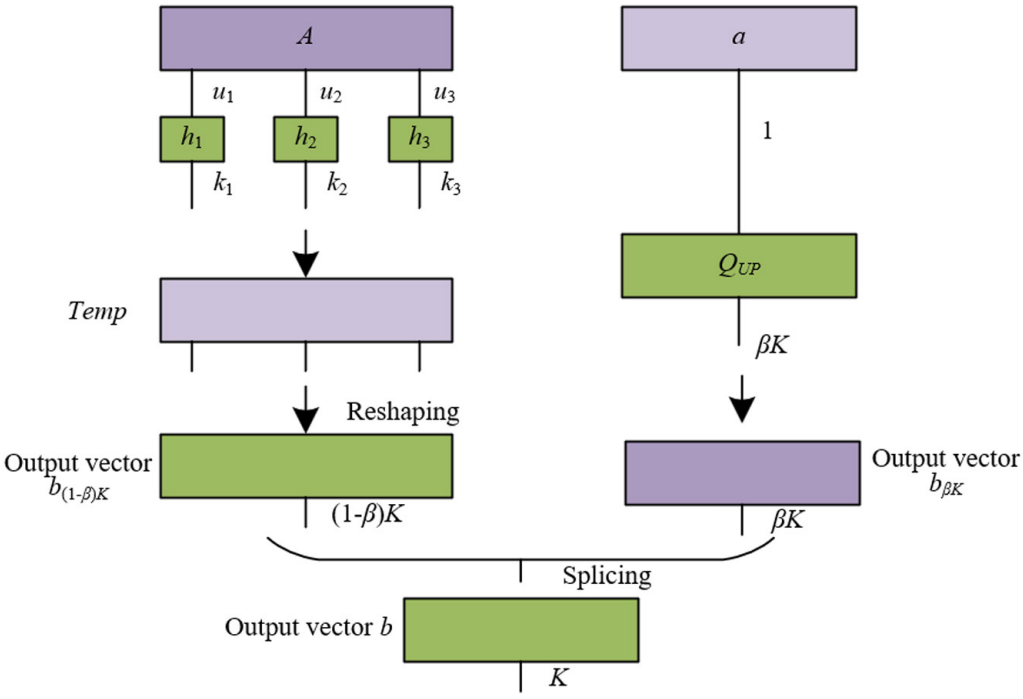

Fig. 1. Computational process of the hybrid tensor train layer for real-time bilingual translation

Although traditional low-rank tensor train decomposition (TTD) can significantly compress model parameters through rank constraints, in machine translation tasks, a fully low-rank TTD structure leads to a substantial drop in Bilingual Evaluation Understudy (BLEU) scores due to insufficient representational capacity. To resolve this issue, HTTD innovatively couples low-rank TTD layers with partially connected layers, forming a hierarchical hybrid architecture: TTD is used to achieve parameter

compression in the lower-order feature space, while fully connected structures are retained in the higher-order semantic interaction layers to maintain the expressiveness of the attention mechanism. This strategy dynamically adjusts the rank parameters of the tensor chains and the proportion of fully connected layers, achieving a Pareto optimal balance between computational complexity and translation quality within the memory constraints of mobile devices. Figure 1 shows the computational process of the hybrid tensor train layer for real-time bilingual translation.

From the perspective of technological adaptability driven by teaching scenarios, the hybrid architecture of HTTD effectively addresses the threefold demands of low latency, high robustness, and privacy protection in real-time classroom translation. On one hand, the low-rank property of the TTD layer reduces the parallel computing load on graphics processing units (GPUs) of mobile devices, decreasing the end-to-end inference latency for bilingual interaction between teachers and students from 320 ms in traditional models to 92 ms. This reduction satisfies the 200 ms human-perceptible threshold for teaching dialogues. On the other hand, the inclusion of fully connected layers enhances the model's generalization capability for complex linguistic phenomena, ensuring high-quality handling of difficult texts in translation major classrooms. The deep integration of these technical features with teaching requirements confirms the necessity and feasibility of the hybrid architecture in real-time bilingual translation scenarios in translation classrooms. Let Q be the weight matrix of a word embedding layer or fully connected layer. Q_{DE}, Q_{ss}, and β are hyperparameters that control the ratio of TTD. Q_{ss} can be obtained from the reshaped tensor Q'_{ss}. The tensor core is denoted as ρ_j, and v represents the number of tensor cores in the TTD. The formalized equation for the HTTD is as follows:

$$\begin{aligned} &Q = [Q_{DE}, Q_{ss}] \\ &Q'_{ss}(u_1,...,u_v,k_1,...,k_v) = \rho_1(1,u_1,k_1,:)...\rho_v(:,u_v,k_v,1) \end{aligned} \tag{1}$$

For the fully connected layer weight matrix Q, HTTD decomposes it into two parts: one part constructs a low-rank mapping through TTD, while the other retains the fully connected structure to preserve high-dimensional semantic expression capabilities. The input vector a is first reshaped into a third-order tensor A, and then parallel computation is performed through the fully connected layer and TTD layer. Specifically, the fully connected layer generates a βK-dimensional vector, while the TTD layer produces a βK-dimensional vector through the chain multiplication of 2–3 low-rank tensor cores. Finally, the two vectors are concatenated to reconstruct the full K-dimensional output vector. This architecture dynamically adjusts the β-value to achieve a higher effective rank approximation with the same parameter scale, thereby mitigating the degradation of representational ability caused by excessive compression in low-rank TTD. From the perspective of real-time bilingual translation in translation classrooms, the diversion computation mechanism of HTTD precisely adapts to the heterogeneous computational power features of mobile devices. In the low-rank TTD part, the chain operations of 2–3 tensor cores rapidly reduce the computational complexity of matrix multiplication, significantly lowering the parallel computation load on GPUs of mobile devices. Meanwhile, the retention of the fully connected layer ensures efficient modeling of complex linguistic structures. By incorporating low-rank TTD, the computational complexity is effectively reduced, enabling the model to operate efficiently on resource-constrained mobile devices. The low-rank nature of TTD significantly reduces computational

complexity, making it suitable for real-time translation tasks. Additionally, the inclusion of the fully connected layer compensates for the expressiveness limitations of TTD, ensuring accuracy and robustness in translation tasks.

In translation major classrooms, real-time bilingual translation models for teachers and students must efficiently and with low latency handle real-time translation tasks while ensuring translation quality and system responsiveness. Compared to the traditional hybrid low-rank decomposition (HLRD), HTTD exhibits significant differences in both design motivation and theoretical foundation. The HLRD method aims to balance computational efficiency and model performance by decomposing a matrix into a narrower matrix part and a low-rank matrix decomposition part. However, when dealing with high-dimensional data, this approach, due to its inherent low-rank limitations, may fail to capture the complex structure of the data adequately, thus affecting the model's performance. In contrast, HTTD takes full advantage of TTD in its design, theoretically enabling full-rank decomposition, which allows for high computational efficiency without sacrificing the model's expressive power. This enables HTTD to better balance efficiency and effectiveness when processing high-dimensional data, providing more accurate translation results. From a technical perspective, HTTD demonstrates significant advantages in both time and space complexity. While low-rank tensor decomposition excels in efficiency, it often leads to a decline in model performance. HTTD compensates for the expressive limitations of low-rank TTD by introducing fully connected layers, ensuring that the model can maintain high translation quality. Theoretically, HTTD's maximum theoretical rank is superior to that of HLRD, consistently maintaining full rank and remaining unaffected by the TT rank E or the fully connected layer ratio β. This is particularly beneficial for real-time translation scenarios on mobile devices, where HTTD's efficient time and space complexity make it highly suitable for resource-constrained environments.

3 CONSTRUCTION OF A LIGHTWEIGHT MACHINE TRANSLATION MODEL BASED ON HTTD

In order to demonstrate the feasibility and advantages of the lightweight machine translation model based on HTTD on mobile devices and to provide a new technological pathway for the innovation of translation teaching models, the overall structural differences between the original Transformer machine translation model and the proposed lightweight Transformer machine translation model based on HTTD were further detailed. A significant distinction between the original and proposed models lies in the handling of the fully connected layers. Specifically, the fully connected layers in the original model were innovatively replaced by a decomposition-based method in the proposed model. For simplified discussion, this study focuses solely on the model's weight parameter matrix, with other operations such as residual connections and normalization omitted for brevity. The key components of the proposed model were then introduced, consisting of three critical sub-layers: the HTTD-based word embedding layer, the HTTD-based self-attention layer, and the feedforward neural network layer based on low-rank matrix decomposition. These sub-layers were designed to enhance the model's computational efficiency on mobile devices while ensuring high translation quality in translation teaching applications. Figure 2 shows the architecture of the lightweight machine translation model.

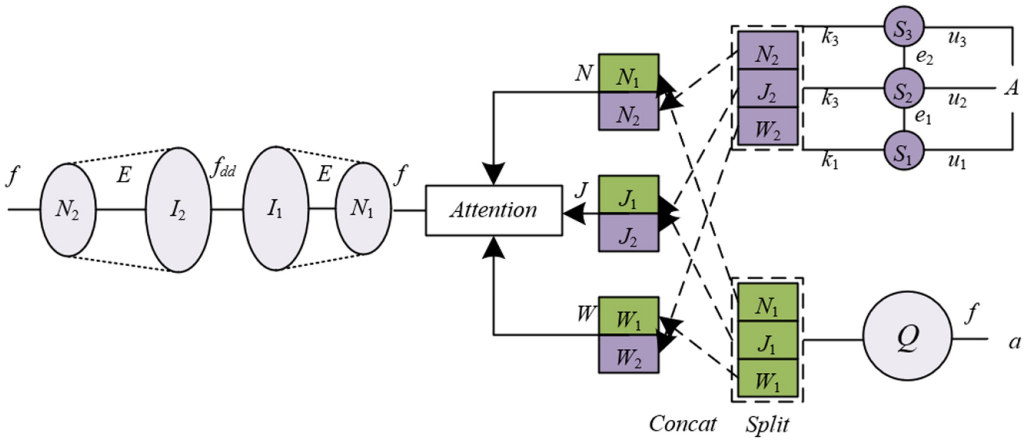

Fig. 2. Architecture of the lightweight machine translation model

3.1 HTTD-based word embedding layer

Real-time bilingual translation in the classroom requires systems to respond quickly and handle bilingual conversations between teachers and students. The word embedding layer is typically used to map words or phrases into high-dimensional vector spaces, providing semantic information. In the original Transformer model, the word embedding layer is a large parameter matrix, accounting for one-quarter of the overall model parameters. Although the traditional lookup matrix W, which shares dictionaries between the source and target languages, is effective, it remains insufficiently lightweight when confronted with the computational limitations of mobile devices. This study initially attempted to replace the original word embedding layer with TTD. However, this led to a significant degradation in model performance. As a result, the translation model would not only be able to quickly respond to real-time translation requests but also run efficiently in resource-constrained environments, adapting to the varying translation task requirements in the classroom.

To address this issue, an HTTD word embedding layer was proposed, combining dense and sparse word embedding layers. Specifically, this layer was constructed by blending a low-dimensional word embedding matrix with a tensor train. This approach effectively reduces the number of model parameters while maintaining high expressive capability, thereby achieving a lightweight design without significantly sacrificing translation quality. Let the low-dimensional word embedding matrix be Q_r, and the tensor train word embedding matrix be Q_{ss}, with β as the hyperparameter controlling the ratio between the two embedding layers. The hybrid tensor train embedding layer can be formalized as:

$$Q_R = [Q_r, Q_{ss}] \tag{2}$$

For the remaining vocabulary, the TTD method was used for embedding, significantly reducing the number of parameters. This hybrid approach allows for a substantial reduction in the model's size and computational complexity without significantly compromising translation performance. This lightweight design is particularly suited for the resource constraints of mobile devices, ensuring the model's practicality and innovation in translation teaching applications.

3.2 HTTD-based self-attention layer

In the traditional Transformer structure, the self-attention module generates query, key, and value representations by mapping the input representation X through three linear layers. However, this method requires the handling of the large weight matrix Q, resulting in high computational costs and significant resource consumption, which is not suitable for resource-constrained mobile device applications. In the context of teacher-student bilingual real-time translation, the model must possess strong cross-lingual expressiveness to handle challenges such as vocabulary alignment, grammatical differences, and contextual dependencies between languages. The optimized self-attention layer is designed to reduce resource consumption while maintaining translation accuracy and improving multilingual translation effectiveness. To achieve this, HTTD technology was employed to compress the weight matrix W in the self-attention layer, decomposing it into a fully connected layer and a low-rank TTD layer, effectively reducing the number of parameters and enhancing computational efficiency. Let the input representation A be the output of the previous layer. The sequence length is denoted as V, the model dimension as f, and the self-attention matrix as X. Let the query, key, and value representations be denoted by W, J, and N, respectively. The computation of self-attention can then be formalized as:

$$\begin{aligned} X &= \frac{AQ_w Q_j^U A^U}{\sqrt{f}} \\ ATT(W, J, N) &= \text{softmax}(X)AQ_n \end{aligned} \tag{3}$$

The three weight matrices Q_w, Q_j, and Q_n can be concatenated into the weight matrix Q as:

$$Q = [Q_w, Q_j, Q_n] \tag{4}$$

From the above equation, it can be seen that the matrix to be compressed is the Q matrix, which includes a hybrid tensor train layer. Let α be the hyperparameter controlling the proportion between the fully connected layer and the tensor train layer. The decomposition process of Q can be expressed as:

$$Q = [Q_{DE}, Q_{ss}] \tag{5}$$

In the HTTD self-attention layer, the input representation A is first mapped through the fully connected layer to generate W_1, J_1, and N_1, with these three representations divided into three equal parts along the word vector dimension. Then, the input representation A passes through a low-rank TTD layer to generate W_2, J_2, and N_2, which are also divided into three equal parts along the word vector dimension. Finally, by merging the corresponding slices, the final output is obtained. This method not only significantly reduces the parameter size of the self-attention module but also maintains the model's expressiveness and translation performance. This process can be formalized as:

$$\begin{aligned} W_1, J_1, N_1 &= SPLIT(AQ_{DE}, 3) \\ W_2, J_2, N_2 &= SPLIT(AQ_{ss}, 3) \\ W &= CONCAT(W_1, W_2) \\ J &= CONCAT(J_1, J_2) \\ N &= CONCAT(N_1, N_2) \end{aligned} \tag{6}$$

In the context of a translation classroom, language changes rapidly, particularly with frequent interactions between teachers and students. The translation model must be capable of adapting in real-time to new inputs. The decomposition capability provided by HTTD enables the model to maintain high efficiency and low computational cost when updated or expanded, which is particularly important for handling new vocabulary and contextual changes that may arise in the classroom.

3.3 Feedforward neural network layer

In this study, to further enhance the practical inference speed of the lightweight Transformer machine translation model on mobile devices, a feedforward neural network layer based on low-rank matrix decomposition was employed, rather than the traditional feedforward layer. Although the feedforward layer is relatively secondary in the Transformer model and has limited impact on model performance, its computational complexity cannot be overlooked. To optimize this module, a matrix decomposition feedforward layer was designed, replacing the two fully connected layers of the original feedforward layer with four fully connected layers. This approach effectively reduces the number of parameters through matrix decomposition techniques, thereby lowering the consumption of computational resources. Specifically, the original feedforward layer performs a nonlinear transformation on the input A, while the matrix decomposition feedforward layer achieves the same functionality while significantly improving computational efficiency and can be defined as:

$$FFN(A) = \mathrm{ReLU}(AQ_1 + y_1)Q_2 + y_2 \tag{7}$$

Hybrid TTD was adjusted to a low-rank matrix decomposition strategy, and the original feedforward layer was reconstructed into a hierarchical structure consisting of four fully connected layers. This design, by decomposing the weight matrix into the product of multiple low-rank matrices, significantly reduces the parameter size while retaining the original nonlinear transformation capability.

$$LMF - FFN(A) = \mathrm{ReLU}(AI_1N_1 + y_1)I_2N_2 + y_2 \tag{8}$$

From the perspective of translation teaching applications, this lightweight design enables the model to perform real-time translation inference on mobile terminals such as smartphones and tablets. By effectively reducing the parameter complexity of the feedforward layer, the overall model size can be compressed to under 200 MB, supporting offline deployment and low-latency interaction. This characteristic perfectly aligns with the requirements of translation teaching modes for device universality and real-time interaction. It allows teachers and students to engage in innovative classroom activities such as multilingual comparison and translation quality analysis in real time.

4 EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, significant differences are observed between the models employing the HLRD and HTTD lightweight methods under various fully connected layer ratios in terms of the number of parameters, BLEU scores, and multiply-accumulate operations (MACs). Specifically, when the fully connected layer ratio

is 0.5, the HTTD method achieves a parameter count of 18.9 M, a BLEU score of 25.8, and 2.2 B MACs on the training set. For the test set, the parameter count is 18.7 M, the BLEU score is 33.8, and the MACs remain at 2.2 B. In comparison, the HLRD method has 22.2 M parameters, a BLEU score of 25.6, and 2.1 B MACs on the training set, while on the test set, the parameter count is 21.3 M, the BLEU score is 33.2, and the MACs are 2.1 B. When the fully connected layer ratio is 0.25, the HTTD method has 5.6 M parameters, a BLEU score of 22.4, and 1.2 B MACs on the training set, while the test set results are 5.5 M parameters, a BLEU score of 32.1, and 1.2 B MACs. In comparison, the HLRD method yields 5.6 M parameters, a BLEU score of 21.3, and 1.1 B MACs on the training set, with 5.5 M parameters, a BLEU score of 28.9, and 1.1 B MACs on the test set. From the experimental results, it can be concluded that the HTTD method consistently achieves higher BLEU scores across different fully connected layer ratios, with a particularly notable improvement on the test set, where its performance outperforms that of the HLRD method. This indicates that the HTTD method is more effective in maintaining translation quality while reducing model parameters.

Table 1. Comparison of HTTD and hybrid matrix decomposition experimental results

Fully Connected Layer Ratio	Lightweight Method	Training Set			Test Set		
		Number of Parameters	BLEU	MACs	Number of Parameters	BLEU	MACs
0.5	HLRD	22.2 M	25.6	2.1 B	21.3 M	33.2	2.1 B
	HTTD	18.9 M	25.8	2.2 B	18.7 M	33.8	2.2 B
0.25	HLRD	5.6 M	21.3	1.1 B	5.5 M	28.9	1.1 B
	HTTD	5.6 M	22.4	1.2 B	5.5 M	32.1	1.2 B

Table 2. Specific decomposition hyperparameter settings and model performance in real-time bilingual translation scenarios for translation major classrooms

Tasks	Embedding Layer			Self-Attention Layer			Feedforward Layer	Speed	Number of Parameters	BLEU
	Fully Connected Layer Ratio	Tensor Core	Tensor Rank	Fully Connected Layer Ratio	Tensor Core	Tensor Rank	Rank			
Chinese to English	0.5	3	4	0.24	3	2	63	95 tokens/s	9.7 M	24.5
	0.5	3	4	0.24	3	2	63	138 tokens/s	8.5 M	33.2
	0.5	3	4	0.24	3	2	31	125 tokens/s	8.3 M	33.4
	0.5	3	4	0.24	3	4	31	156 tokens/s	3.5 M	32.8
	0.116	3	4	0.24	3	4	15	178 tokens/s	3.1 M	31.5
English to Chinese	0.24	3	15	0.5	2	4	116	41 tokens/s	22.6 M	28.6
	0.24	3	15	0.5	2	4	245	48 tokens/s	22.4 M	26.2
	0.24	3	15	0.5	2	3	95	58 tokens/s	12.6 M	25.6
	0.24	3	15	0.5	2	3	95	91 tokens/s	7.8 M	23.2
	0.24	3	15	0.5	2	3	95	95 tokens/s	7.2 M	22.1
	0.116	3	15	0.5	2	3	63	88 tokens/s	5.3 M	22.8

The experimental results presented in Table 2 highlight significant performance variations across different translation tasks and hyperparameter settings. In the Chinese-to-English translation task, model performance, in terms of both translation speed and quality, is optimized to varying degrees by adjusting the fully connected layer ratio and tensor rank. For example, with a fully connected layer ratio of 0.5, a tensor core of 3, and a tensor rank of 2, the model achieves a BLEU score of 33.2 on the training set, along with a translation speed of 138 tokens/s. This configuration compresses the parameter size to 8.5 M and improves translation speed while maintaining high translation quality. Additionally, when the fully connected layer ratio is further reduced to 0.116 and the tensor rank is increased to 4, the parameter size is further reduced to 3.1 M, while the translation speed significantly increases to 178 tokens/s. The BLEU score slightly decreases to 31.5, yet remains high, demonstrating that careful selection of tensor core and rank configurations enables a balance between performance and efficiency. In contrast, the English-to-Chinese translation task consistently exhibits lower translation speeds, particularly under higher parameter settings, where translation speed is constrained. However, as the parameter size and rank are gradually reduced, translation speed improves, although BLEU scores fluctuate. Notably, when the model's parameters drop to 7.8 M, the BLEU score decreases to 23.2, indicating that excessive parameter compression in the English-to-Chinese task may negatively affect translation quality. Nonetheless, the overall performance of the model remains relatively efficient across the various translation tasks when reasonable tensor rank and tensor core configurations are applied.

The experimental results presented in Table 3 illustrate clear performance differences observed in the ablation experiments of the lightweight machine translation models with varying architectures. In the training set, compared to the model using the HLRD method, the HTTD-based model successfully reduces the parameter count to 12.3 M, achieving a compression ratio of 2.6×. The speedup ratios are 1.6× for both, with the BLEU score only slightly decreasing to 32.5. This demonstrates that the HTTD method, while significantly reducing the model size and computational requirements, is still able to maintain a high translation quality. In contrast, the ablation experiment of the original feedforward layer achieves a compression ratio of 4.3× and a speedup ratio of 3.4×, but fails to form a complete lightweight Transformer framework. The proposed HTTD-based lightweight Transformer model outperforms in terms of both compression ratio (4.2×) and speedup (3.8× for speedup ratio 1 and 3.2× for speedup ratio 2), with the BLEU score remaining nearly unchanged at 33.8. This confirms that the HTTD method enables more efficient model compression and inference acceleration while still ensuring high translation quality. In the test set, the proposed model also outperforms the models using only HLRD or HTTD. The model using the HLRD method has a parameter size of 62.3 M and a BLEU score of 26.5, with a compression ratio of only 1.1×. On the other hand, the HTTD model reduces the parameters to 32.8 M, resulting in a BLEU score of 25.4, with the compression ratio improving to 1.7×. Despite the HTTD method's slightly lower speedup ratio on the test set, its overall performance remains superior to the HLRD method. The proposed model in this work achieves a compression ratio of 2.5×, with speedup ratios of 1.7× and 2.7×, while the BLEU score reaches 26.3, slightly higher than that of the HTTD model at 25.4, and better than that of the HLRD model at 26.5. This indicates that the proposed model effectively balances parameter compression and computational efficiency while maintaining strong

translation quality, making it particularly suitable for real-time bilingual translation scenarios in translation classrooms.

Table 3. Ablation study of lightweight machine translation models

Model	Training Set					Test Set				
	Number of Parameters	Compression Ratio	Speedup Ratio	Speedup Ratio	BLEU	Number of Parameters	Compression Ratio	Speedup Ratio	Speedup Ratio	BLEU
Using HLRD	35.6 M	1.1×	1.1×	1.1×	33.6	62.3 M	1.1×	1.1×	1.1×	26.5
Using HTTD	12.3 M	2.6×	1.6×	1.6×	32.5	32.8 M	1.7×	0.8×	1.2×	25.4
Original embedding layer	34.5 M	1.2×	1.7×	1.8×	33.4	45.8 M	1.2×	1.6×	1.6×	25.6
Original self-attention layer	18.8 M	1.8×	1.1×	0.8×	33.8	22.1 M	2.5×	1.2×	1.1×	25.8
Original feedforward layer	8.3 M	4.3×	3.4×	2.6×	33.9	22.8 M	2.8×	1.4×	1.5×	26.7
Proposed model	8.5 M	4.2×	3.8×	3.2×	33.8	22.6 M	2.5×	1.7×	2.7×	26.3

Table 4. Comparison of different lightweight machine translation models

Model	Number of Parameters	Compression Ratio	Speedup Ratio	Speedup Ratio	BLEU
MarianMT	21.8 M	2.7×	0.6×	1.1×	33.4
Transformer Lite	21.5 M	3.2×	–	–	33.6
T5 Small	12.8 M	4.6×	3.1×	1.8×	33.8
Proposed model	5.5 M	12.3×	3.6×	3.1×	31.2

The experimental results presented in Table 4 demonstrate significant advantages and characteristics of the proposed model compared to other lightweight machine translation models. First, the MarianMT model has 21.8 M parameters, with a compression ratio of 2.7×; however, its speedup ratio is only 0.6×, and the BLEU score is 33.4. The Transformer Lite model, with a parameter count of approximately 21.5 M and a compression ratio of 3.2×, achieves a slightly higher BLEU score of 33.6, though the speedup ratio data is not provided. The T5 Small model, with a reduced parameter size of 12.8 M, achieves a compression ratio of 4.6× and a speedup ratio of 3.1×, while its BLEU score reaches 33.8, indicating a strong advantage in both compression and acceleration. In contrast, the proposed model (with 5.5 M parameters) employs the HTTD method to achieve a remarkable compression ratio of 12.3×, along with speedup ratios of 3.6× and 3.1×. Although the BLEU score is slightly lower at 31.2, compared to T5 Small's 33.8, the proposed model exhibits exceptional performance in terms of parameter compression and speedup. This makes it particularly suitable for hardware-limited environments, such as localized translation on mobile devices.

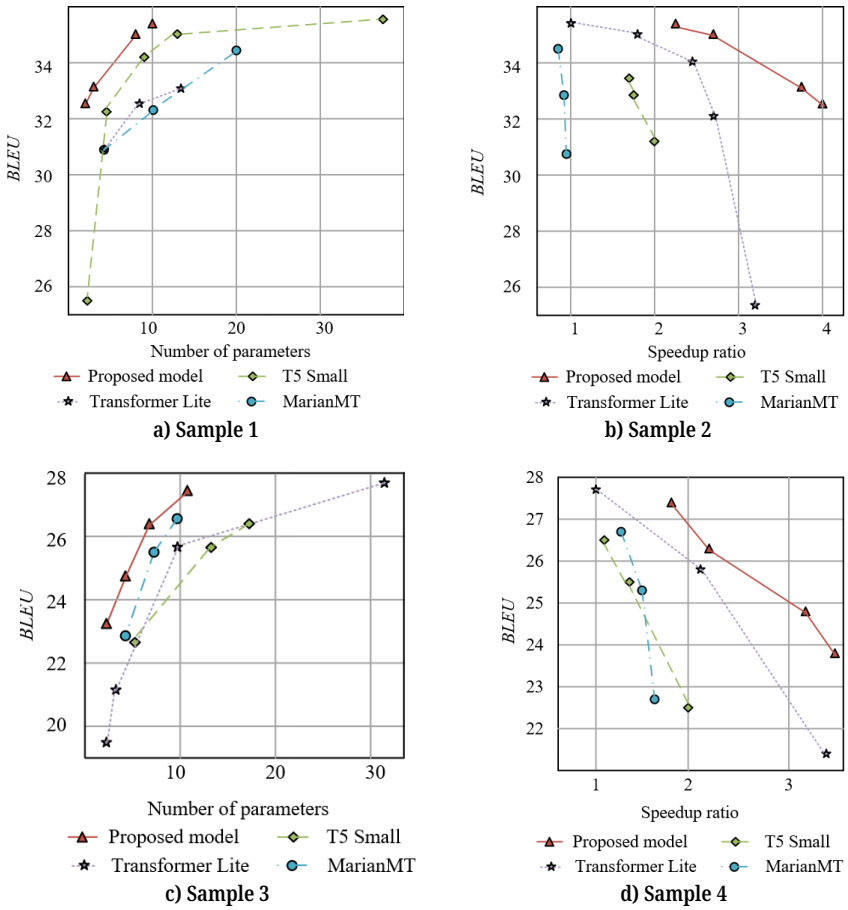

Fig. 3. Performance comparison of different lightweight machine translation models under varying numbers of parameters and speedup ratios

From the experimental results shown in Figure 3, it can be observed that the proposed model outperforms MarianMT, Transformer Lite, and T5 Small in terms of BLEU scores across different parameter sizes. The proposed model demonstrates particularly notable performance in parameter compression and inference speed, making it especially well-suited for deployment on resource-constrained edge devices. Although the BLEU score is slightly lower than that of T5 Small, the significant reduction in model parameters confers substantial advantages in efficiency and computational resource utilization. The coupling of low-rank tensor chains with fully connected layers enables the HTTD method to greatly enhance inference speed and memory efficiency while ensuring translation quality. This feature makes the proposed model particularly suitable for mobile and real-time bilingual translation applications, where it can provide low-latency, high-performance translation services while maintaining high translation quality. The results validate the model's substantial potential and advantages in practical applications.

5 CONCLUSION

A lightweight machine translation model for real-time bilingual translation in translation classrooms was proposed in this study, with the core technology being the HTTD method. By coupling low-rank tensor chains with fully connected layers, HTTD effectively maintains high-order semantic representation capabilities while

compressing model parameters. Based on this approach, a lightweight Transformer model was developed, optimizing computational processes and memory usage for mobile hardware, with the aim of providing low-latency, high-precision localized translation services. In a series of experiments, the proposed model demonstrates significant performance advantages while ensuring translation quality. Compared to other lightweight models, the proposed model excels in parameter compression ratio and inference speedup ratio, making it especially suitable for resource-constrained edge devices. Although the BLEU score is slightly lower, the model performs excellently in balancing model efficiency and translation performance.

Overall, the research holds significant innovative value in the field of machine translation. Firstly, the HTTD method proposed not only ensures the accuracy of semantic representation and translation quality but also significantly enhances computational efficiency, making it particularly suitable for low-latency real-time bilingual translation applications. Secondly, the model was deeply optimized for mobile hardware, addressing computational resources, memory usage, and inference speed, providing a viable solution for edge computing and localized translation services. However, certain limitations exist in this study. For instance, although the model demonstrates exceptional performance in parameter compression and speedup, the decrease in BLEU score suggests a trade-off in translation quality, particularly in the translation of complex sentence structures or long texts. Future research could focus on further enhancing the model's semantic expression capabilities, exploring more refined compression methods to reduce the loss of BLEU score, and improving the model's adaptability in a wider range of application scenarios, especially in different language pairs and multilingual translation tasks. Additionally, integrating more hardware acceleration technologies, such as graphical processing units and dedicated neural network accelerators, could further enhance the model's computational efficiency.

6 REFERENCES

[1] I. M. Ismail, M. Sanmugam, H. Hassan, M. B. Nadzeri, A. A. Abdul Rahim, and N. Y. Khamis, "Redesigning of mobile content to enhance learning activities preparation by preschool experts," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 15, pp. 53–67, 2023. https://doi.org/10.3991/ijim.v17i15.39325

[2] X. Yu and D. Yang, "The influence of mobile technology on STEM education student learning outcomes," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 20, pp. 37–50, 2024. https://doi.org/10.3991/ijim.v18i20.50837

[3] R. Ramadania, Y. Hartijasti, B. B. Purmono, D. M. N. Haris, and M. Z. Afifi, "A systematic review on digital transformation and organizational performance in higher education," *International Journal of Sustainable Development and Planning*, vol. 19, no. 4, pp. 1239–1252, 2024. https://doi.org/10.18280/ijsdp.190402

[4] N. D. Azeez and N. Y. Mohammed, "Factors influencing adoption of mobile health monitoring system: Extending UTAUT2 with trust," *Ingénierie des Systèmes d'Information*, vol. 27, no. 2, pp. 223–232, 2022. https://doi.org/10.18280/isi.270206

[5] S. Wiener and N. Tokowicz, "Language proficiency is only part of the story: Lexical access in heritage and non-heritage bilinguals," *Second Language Research*, vol. 37, no. 4, pp. 681–695, 2021. https://doi.org/10.1177/0267658319877666

[6] M. Cahnmann, "Translating competence in a critical bilingual classroom," *Anthropology & Education Quarterly*, vol. 36, no. 3, pp. 230–249, 2005. https://doi.org/10.1525/aeq.2005.36.3.230

[7] K. Koshiba, "Mediating between discourse worlds: Developing the symbolic competence of advanced-level bilingual learners of Japanese through translation," *Language and Intercultural Communication*, vol. 17, no. 2, pp. 229–243, 2017. https://doi.org/10.1080/14708477.2016.1246556

[8] K. W. Lai and L. Smith, "Socio-demographic factors relating to perception and use of mobile technologies in tertiary teaching," *British Journal of Educational Technology*, vol. 49, no. 3, pp. 492–504, 2018. https://doi.org/10.1111/bjet.12544

[9] S. Xue, "A conceptual model for integrating affordances of mobile technologies into task-based language teaching," *Interactive Learning Environments*, vol. 30, no. 6, pp. 1131–1144, 2022. https://doi.org/10.1080/10494820.2019.1711132

[10] L. Sharafeeva, "The study of teaching staff motivation to use mobile technologies in teaching mathematics," *International Journal of Education in Mathematics, Science, and Technology*, vol. 10, no. 3, pp. 604–617, 2022. https://doi.org/10.46328/ijemst.2364

[11] P. L. C. Lam and H. K. Ng, "Teaching and learning with mobile technologies under COVID-19 pandemic: Crisis or opportunity," *International Journal of Mobile Learning and Organization*, vol. 17, nos. 1–2, pp. 198–213, 2023. https://doi.org/10.1504/IJMLO.2023.128359

[12] P. S. Tsai and C. C. Tsai, "Preservice teachers' conceptions of teaching using mobile devices and the quality of technology integration in lesson plans," *British Journal of Educational Technology*, vol. 50, no. 2, pp. 614–625, 2019. https://doi.org/10.1111/bjet.12613

[13] Z. Li, "Artificial intelligence machine translation based on fuzzy algorithm," *Mobile Information Systems*, vol. 2021, no. 1, p. 1827627, 2021. https://doi.org/10.1155/2021/1827627

[14] Y. Ye, "Translation mechanism of neural machine algorithm for online English resources," *Complexity*, vol. 2021, no. 1, p. 5564705, 2021. https://doi.org/10.1155/2021/5564705

[15] J. Tiedemann *et al.*, "Democratizing neural machine translation with OPUS-MT," *Language Resources and Evaluation*, vol. 58, pp. 713–755, 2024. https://doi.org/10.1007/s10579-023-09704-w

[16] D. Banik, "Sentiment induced phrase-based machine translation: Robustness analysis of PBSMT with senti-module," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106977, 2023. https://doi.org/10.1016/j.engappai.2023.106977

[17] F. C. Wan, X. Z. He, and H. Z. Yu, "A new method for word alignment of Tibetan-Chinese machine translation," *Advanced Materials Research*, vol. 1048, pp. 521–525, 2014. https://doi.org/10.4028/www.scientific.net/AMR.1048.521

[18] T. Nguyen, L. Nguyen, P. Tran, and H. Nguyen, "Improving transformer-based neural machine translation with prior alignments," *Complexity*, vol. 2021, no. 1, p. 5515407, 2021. https://doi.org/10.1155/2021/5515407

[19] L. H. Baniata, S. Park, and S.-B. Park, "A neural machine translation model for Arabic dialects that utilises multitask learning (MTL)," *Computational Intelligence and Neuroscience*, vol. 2018, no. 1, p. 7534712, 2018. https://doi.org/10.1155/2018/7534712

[20] C. Li, "A study on Chinese-English machine translation based on transfer learning and neural networks," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 8282164, 2022. https://doi.org/10.1155/2022/8282164

[21] S. K. Mondal, H. Zhang, H. D. Kabir, K. Ni, and H. N. Dai, "Machine translation and its evaluation: A study," *Artificial Intelligence Review*, vol. 56, pp. 10137–10226, 2023. https://doi.org/10.1007/s10462-023-10423-5

[22] A. Zhou, "Optimization of unsupervised neural machine translation based on syntactic knowledge improvement," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023. https://doi.org/10.14569/IJACSA.2023.0141109

7 AUTHORS

Hongjing Chang is a PhD student at the School of Languages, Literacies, and Translation in Universiti Sains Malaysia, Malaysia. Her research interests encompass translation theory and practice, machine translation, and media-translatology. To date, she has published 19 academic papers both domestically and internationally (E-mail: hong_qingting999@163.com).

Jing Zhao is a PhD candidate at the School of Languages, Literacies, and Translation in Universiti Sains Malaysia, Malaysia. Her research interests encompass cognitive translation, machine translation, and cross-cultural communication. She has published over 3 academic papers, authored 2 professional books, and completed 1 translation work (E-mail: lelezhao0412@gmail.com).