scientific reports

Check for updates

OPEN

# The analysis of learning investment effect for artificial intelligence English translation model based on deep neural network

Yan Zhang[1] & Shuangshuang Lyu[2]✉

**With the rapid development of multimodal learning technologies, this work proposes a Future-Aware Multimodal Consistency Translation (FACT) model. This model incorporates future information guidance and multimodal consistency modeling to improve translation quality and enhance language learning efficiency. The model innovatively integrates target future contextual information with a multimodal consistency loss function, effectively capturing the interaction between text and visual information to optimize translation performance. Experimental results show that, in the English-German translation task, the FACT model outperforms the baseline model in both Bilingual Evaluation Understudy (BLEU) and Meteor scores. The model achieves BLEU scores of 41.3, 32.8, and 29.6, and Meteor scores of 58.1, 52.6, and 49.6 on the Multi30K tset16, tset17, and Microsoft Common Objects in Context datasets, respectively, demonstrating its remarkable performance advantages. Significance analysis also verifies this result. Ablation experiments indicate that the future context information supervision function and multimodal consistency loss function are crucial for the model's performance. Further language learning experiments show that the FACT model significantly outperforms the Transformer model in multiple metrics, encompassing learning efficiency (83.2 words/hour) and translation quality (82.7 points), illustrating its potential in language learning applications. In short, the FACT model holds high application value in multimodal machine translation and language learning. This work provides new ideas and methods, and advances future multimodal translation technology research and applications.**

**Keywords** Multimodal translation, Future information guidance, Neural networks, Learning investment effect, Visual information

## Research background and motivations

With the swift progress of Artificial Intelligence (AI) technology, Neural Machine Translation (NMT) has achieved significant breakthroughs in language translation. Translation models based on Deep Neural Network (DNN) can capture sentence-level contextual information and realize an end-to-end translation framework. Compared to traditional Statistical Machine Translation (SMT), these models offer substantial advantages in translation quality and fluency[1–3]. However, existing NMT models still face many limitations, such as poor performance in handling polysemous words, translating long sentences, and lacking contextual semantic consistency[4–6]. Moreover, with the advancement of educational technologies, the potential of translation models in assisting language learning has gradually garnered attention. However, research on how these models can enhance learners' language abilities and learning investment effects remains limited[7–9].

In recent years, the introduction of multimodal technology has provided new opportunities for development in NMT. Multimodal translation models can better capture semantic relationships in complex contexts by integrating visual and linguistic information, particularly excelling in descriptive text translation tasks[10,11]. Meanwhile, attention mechanisms, as one of the core techniques in deep learning, can effectively capture long-distance dependencies between sentences, offering theoretical support for further optimizing multimodal translation models[3,12,13]. However, utilizing multimodal features and attention mechanisms can improve the translation models' performance. Applying them in language learning scenarios to enhance learners' experiences and outcomes remains a question worthy of deeper exploration. Based on this, this work proposes a translation

[1]School of International Education, Jilin Engineering Normal University, Changchun 130052, China. [2]Shenyang Institute of Engineering, Shenyang 110137, China. ✉email: zhangyan1219@jlenu.edu.cn

model based on multimodal consistency and future target context modeling, named the Future-Aware Multimodal Consistency Translation (FACT) model. The work conducts experimental validation to assess the model's effectiveness in assisting language learning, focusing on its impact on users' learning investment.

## Research objectives

This work aims to design and validate a DNN-based multimodal consistency translation model while exploring its potential in improving translation performance and learning investment effects. Specifically, a future target context capture method based on the attention mechanism is proposed, integrating visual and textual features to enhance the translation model's semantic consistency and contextual understanding ability. Meanwhile, the performance of the FACT model is experimentally validated across diverse datasets, sentence lengths, and translation scenarios, and the role of visual features in improving translation quality is analyzed. Moreover, the application of the model in language learning scenarios is explored, and its impact on users' learning investment effects is evaluated, providing theoretical support and practical reference for the educational technology field. The main contributions of this work include the following aspects. It proposes an NMT framework that combines multimodal consistency mechanisms with future target context modeling, remarkably enhancing the model's semantic capture and contextual understanding abilities. An innovative future context supervision mechanism based on multimodal information is designed to effectively address the shortcomings of traditional models in translating long sentences and polysemous words. Through extensive ablation experiments and cross-lingual validation, the FACT model's superior performance and generalization ability in multimodal machine translation (MMT) tasks are systematically demonstrated. The practical application of the multimodal translation model in assisting language learning is explored, evaluating its positive impact on learners' language proficiency improvement and learning engagement, thus expanding the application prospects of multimodal translation technologies in education. This work lays the foundation for developing AI translation technologies and applying intelligent language learning by constructing an innovative translation model.

## Literature review

Over the years, NMT technology has been widely applied across multiple domains, including language learning and natural language processing (NLP). Zhang (2024) investigated a recurrent neural network (RNN)-based English speech translation system, examining its advantages in processing time-series data and validating its strong performance in accuracy, fluency, and real-time capabilities[14]. Regarding machine learning applications in language education, Klimova et al. (2023) demonstrated that NMT effectively developed language skills, encompassing speaking, writing, reading, listening, and mediation, particularly benefiting advanced L2 learners. Their research indicated that NMT could serve as a powerful online reference tool when teachers properly introduced its capabilities and limitations to students for optimal pedagogical outcomes[15]. Ohashi (2025) employed mixed methods to explore language teachers' perceptions and practices regarding machine translation (MT). This research found that while most educators held positive attitudes toward personal use, they generally lacked instructional guidance and training, with many expressing a need for additional support[16].

With the continuous evolution of multimodal learning technologies, increasing academic attention has focused on integrating visual, auditory, and textual modalities to enhance model performance. Zhang et al. (2024) developed multiple end-to-end lower-limb motion recognition frameworks using deep learning and multimodal information, markedly improving recognition accuracy for both healthy subjects and stroke patients[17]. Tomar et al. (2024) proposed an innovative emotion recognition approach combining facial expressions and vocal cues through feature-level and decision-level fusion techniques, validating multimodal integration's potential for complex semantic tasks[18]. As an important branch of multimodal research, multimodal neural machine translation (MNMT) has achieved numerous advancements in recent years. Guo et al. (2024) proposed a multimodal fusion strategy guided by a multi-granularity visual pivot, which eliminated language differences between languages through cross-modal contrastive decoupling, thus effectively improving the performance of MNMT[19]. Li et al. (2023) introduced a semi-supervised learning-based multimodal information fusion method that trained a multimodal attention network with a small parallel corpus. This method integrated text and image information to achieve more precise automated MT[20]. Zhao et al. (2022) presented a region-attention MNMT method based on semantic image regions. This method combined semantic image regions extracted through object detection with text features and used RNNs and self-attention network architectures, thereby enhancing translation quality and accuracy[21]. Analysis of these studies reveals that multimodal mechanisms effectively enhance task performance primarily through extracting complementary semantic information from different modalities. The visual modality strengthens perception of spatial, entity, and verb semantics in text, alleviating linguistic ambiguity, unclear references, and contextual dependency issues. The auditory modality supplements rhythmic, emotional, and tonal cues. Multimodal fusion techniques, including attention mechanisms and contrastive learning, enable selective focus on critical modality signals and cross-modal semantic alignment, thereby improving model generalization and robustness in semantic modeling. Particularly in language learning applications, multimodal information increases translation accuracy and provides intuitive support during learners' perception of linguistic input, effectively facilitating language comprehension and vocabulary retention.

Although existing MNMT methods have made progress in improving translation quality, current research still has shortcomings in the following aspects. Insufficient capability in consistent modeling of multimodal features may lead to information redundancy or interference. The lack of practical evaluation oriented to language learning tasks makes it difficult to measure the application value of multimodal translation in educational contexts. Future context information has not been effectively integrated, limiting the model's performance in complex language environments. To address these shortcomings, this work proposes a FACT model incorporating future information guidance and multimodal consistency modeling. This model effectively

utilizes image and text features to improve translation quality and strengthen its generalization ability by introducing target future context information and a multimodal consistency loss function.

## Research model

### Future information-guided MMT model

The goal of MMT is to generate a target sentence $\{y_1, y_2, \ldots, y_m\}$ given an image *pic* and a source sentence $\{x_1, x_2, \ldots, x_n\}$. $n$ and $m$ represent the lengths of the source and target sentences, respectively. The proposed FACT model is built on the traditional Transformer architecture. It unifies the source sentence and visual features into a shared semantic representation $s_{enc}$ through the encoder. In this process, the multi-head attention mechanism extracts the global contextual information of the source sentence and generates hidden state vectors that capture the semantics of the source sentence. The decoder utilizes the semantic information in $s_{enc}$ to generate the current word $y_t$ of the target sequence at each time step $t$. Figure 1 illustrates the FACT model's overall process.

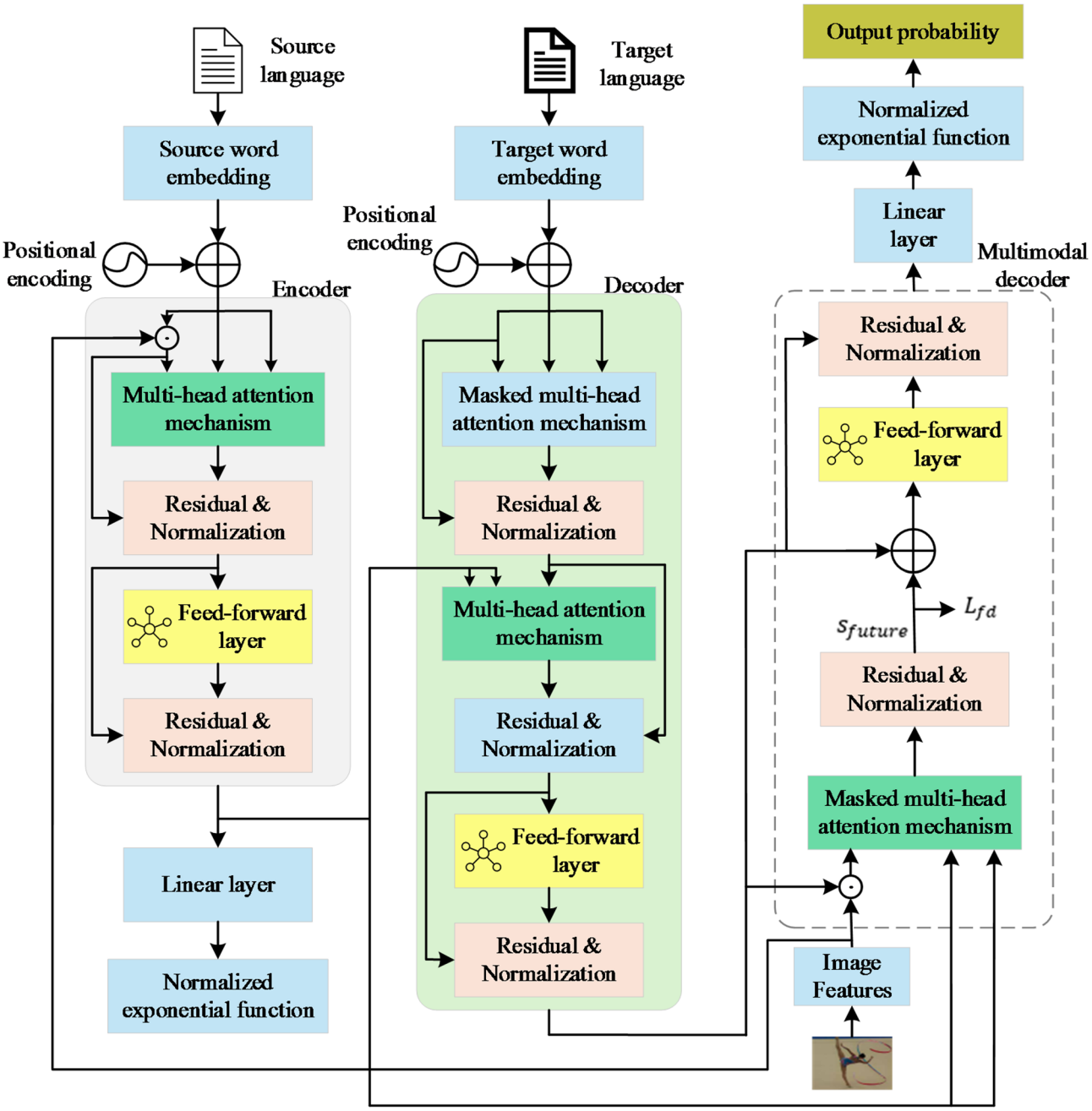


**Fig. 1**. Flowchart of the future information-guided MMT model.

The FACT model's encoder processes the source sentence and image inputs through a shared architecture. This integrated approach simultaneously extracts and combines visual and linguistic features to predict future contextual information at the encoder output, thus achieving a consistent MMT process.

### Future target context capture method based on attention mechanism

Incorporating future contextual information into the decoder can substantially enhance the translation model's performance. However, traditional text translation models are limited by the inability of the decoder to observe future semantic information[1,22,23]. Based on this, a future target context prediction mechanism utilizing visual features is proposed, which captures future context representations through visual information modeling. The process is presented in Fig. 2.

First, ResNet50 is employed to extract features from the image *pic*, generating a visual feature vector $v_{pic}$:

$$v_{image} = CNN_{image}(pic) \tag{1}$$

$v_{image}$ contains semantic information consistent with the text representation. Next, the encoder output $s_{enc} \in \mathbb{R}^{N\times d}$, the decoder output $s_{dec} \in \mathbb{R}^{N\times d}$, and the visual feature $v_{image} \in \mathbb{R}^{M\times d}$ are concatenated into $q \in \mathbb{R}^{(N+M)\times d}$, which serves as the input to the multi-head attention module:

$$q = [s_{enc}, v_{image}] \tag{2}$$

$N$ and $M$ represent the number of text and image samples, respectively; $d$ refers to the model dimension. The future target context representation $s_{future,t}$ at time step $t$ is calculated through the multi-head attention mechanism as follows:

$$s_{future,t} = \sum_{j=1}^{n} \alpha_{i,j}(s_{dec,j}W_V) \tag{3}$$

$\alpha_{i,j}$ denotes the attention weight, which is defined as:

$$\alpha_{i,j} = softmax(\frac{(s_{dec,j}W_K)\cdot(qW_Q)^T}{\sqrt{d}}) \tag{4}$$

$W_K$, $W_Q$, and $W_V$ represent the key, query, and value matrices, respectively.

Additionally, a bag-of-words (BoW) constraint mechanism is introduced, which uses the decoder's hidden state $s_{future,t}$ to predict the future target words $(y_{t+1}.y_{t+2}\ldots y_m)$:

$$P_{BoW}\left(y_{t+1}.y_{t+2}\ldots y_m|s_{future,t}\right) = \prod_{k=t+1}^{m} P(y_k|s_{future,t}) \tag{5}$$

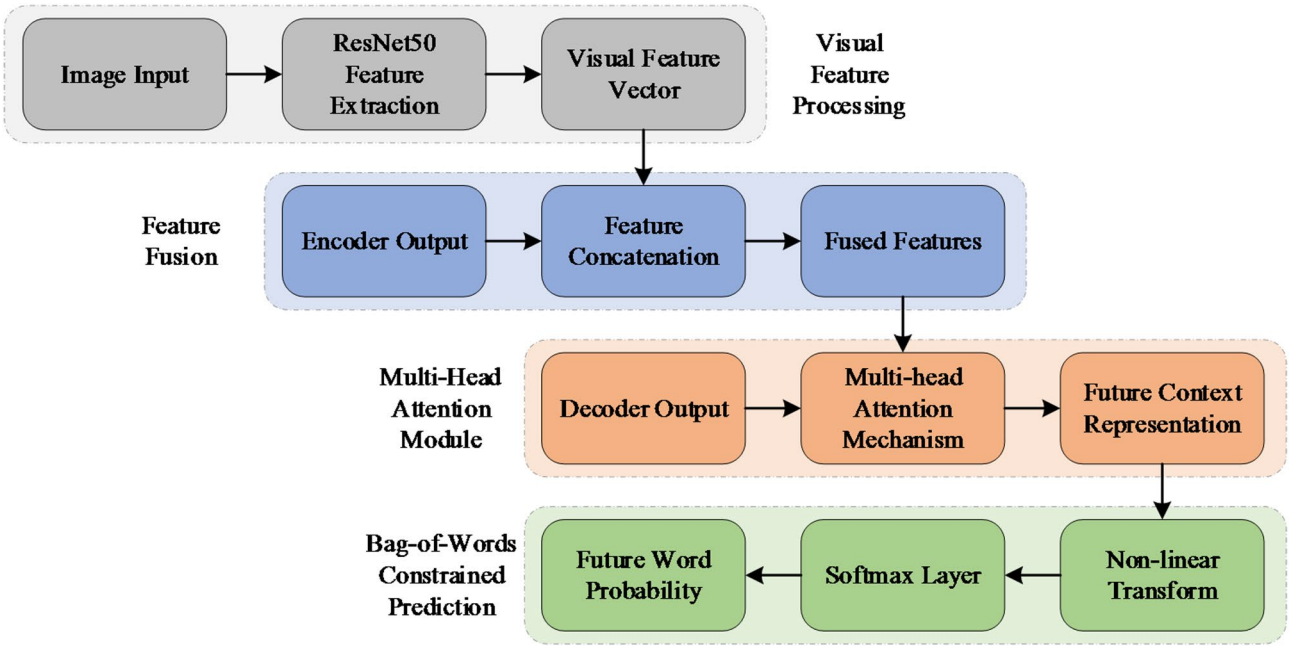


**Fig. 2**. The process of the future target context capture method based on the attention mechanism.

$$P\left(y_k|s_{future,t}\right) = f_s(f_t(s_{future,t})) \tag{6}$$

$P_{BoW}$ represents the predicted probability of future target words based on the BoW constraint; $P\left(y_k|s_{future,t}\right)$ denotes the predicted probability of the target word $y_k$ given the hidden state. $f_s$ means the Softmax layer; $f_t$ stands for the nonlinear transformation layer. Through this mechanism, the future target context can effectively capture the BoW information on the target side, thus improving translation quality.

## Multimodal consistency modeling method

A multimodal consistency modeling method is designed to maximize the utilization of future target context information provided by images. This method enables the model to extract image-based semantic information by enforcing consistency between linguistic and visual representations, thus enhancing MNMT system performance through effective fusion of textual and visual features. As a multimodal input, the visual feature representation of the image needs to contain the same semantic information as the source and target sentences. However, the image feature representation process faces challenges from irrelevant noise regions that may adversely affect model accuracy[24–26]. Therefore, by introducing an image-aware attention mechanism, the model uses textual information as a guide to extract regions from the image that are semantically consistent with the text. Specifically, the source sentence $x$ and visual feature $v_{pic}$ are concatenated and input into the self-attention module to compute the hidden state $h_{enc}$:

$$h_{enc} = \sum_{j=1}^{n} \alpha_{i,j}(x_j W_V) \tag{7}$$

Here, the attention weight $\alpha_{i,j}$ is defined as:

$$\alpha_{i,j} = softmax(\frac{(x_j W_K) \cdot (IW_Q)^T}{\sqrt{d}}) \tag{8}$$

The concatenated input $I$ is represented as:

$$I = [x, v_{pic}] \in \mathbb{R}^{(N+M)\times d} \tag{9}$$

Based on the consistency model, the encoder output $h_{enc}$ is employed for predicting the target sentence, and the probability distribution is expressed as:

$$P_{MD}\left(y_1.y_2 \ldots y_m|h_{enc}\right) = \prod_{k=1}^{m} P(y_k|h_{enc}) \tag{10}$$

$$P\left(y_k|h_{enc}\right) = f_s(f_t(h_{enc})) \tag{11}$$

$f_t$ refers to the nonlinear layer; $f_s$ represents the Softmax layer.

The future context information $s_{future}$ and the decoder state $s_{dec}$ are concatenated and input into the feedforward neural network to generate a more accurate translation representation:

$$U = FeedForward([s_{future}, s_{dec}]) \tag{12}$$

Finally, the objective function for the translation probability is calculated as:

$$P\left(y_k|x, v_{pic}, t\right) = softmax(U_t) \tag{13}$$

To optimize the multimodal translation model, a joint loss function is employed, which includes the text translation loss $L_t$, future target context loss $L_{fd}$, and multimodal consistency loss $L_{md}$:

$$L = aigmin(L_t + L_{fd} + L_{md}) \tag{14}$$

$$L_t = \frac{1}{m}\sum_{t=1}^{m} \log P(y_t|y < t; x; v) \tag{15}$$

$$L_{fd} = \frac{1}{m-1}\sum_{t=1}^{m-1} \log P_{fd}((y_{t+1}, \ldots, y_m)|s_{future,t}) \tag{16}$$

$$L_{md} = \log P_{md}((y_1, y_2, \ldots, y_m)|h_{enc}) \tag{17}$$

$P(y_t|y < t; x; v)$ refers to the predicted probability of the target word $y_t$ given the target word sequence $y < t$, the source sentence x, image $v$ before the current time step. $P_{fd}((y_{t+1}, \ldots, y_m)|s_{future,t})$ represents the predicted probability of the target word sequence $(y_{t+1}, \ldots, y_m)$ from time step t+1 to m, given the future

target context $s_{future,t}$. $P_{md}((y_1, y_2, \ldots, y_m)|h_{enc})$ represents the predicted probability of the target sentence $(y_1, y_2, \ldots, y_m)$ given the encoder output $h_{enc}$.

During the training phase, the model optimizes the parameters of $h_{enc}$ and $s_{future}$ based on the supervised signal from the target sentence. In the inference phase, the trained parameters are used to generate the translation results.

## Experimental design and performance evaluation

### Datasets collection

This experiment employs two widely recognized standard datasets in MMT: Multi30K and Microsoft Common Objects in Context (MS COCO)[27,28]. The Multi30K dataset comprises image-text pairs spanning various domains and is commonly used for image caption generation and multimodal translation tasks. The dataset contains three language pairs: English to German (En-De), English to French (En-Fr), and English to Czech (En-Cs). Specifically, the Multi30K training set encompasses 29,000 bilingual parallel sentence pairs, 1000 validation samples, and 1000 test samples. Each sentence is paired with an image to ensure the consistency between the text description and the image content, thus providing high-quality multimodal data for model training. The test16 and test17 datasets are used here. MS COCO is a dataset containing a wide range of images and their descriptions, extensively used in multiple tasks in computer vision and NLP. Beyond its established role as a standard benchmark for image captioning evaluation, the dataset's rich semantic annotations make it particularly suitable for assessing model performance in cross-domain and cross-lingual translation scenarios.

### Experimental environment

This experiment utilizes the Fairseq toolkit built upon the PyTorch framework. Fairseq is an open-source toolkit widely used in NLP tasks, particularly for constructing and training MT models. It supports various model architectures, including RNNs, convolutional neural networks, and Transformers, enabling effective performance enhancement in MT tasks. Based on Fairseq, the experimental model framework can be easily constructed, and the corresponding training tasks can be configured. The toolkit provides efficient parallel computing support and optimized training workflows, enabling effective large-scale model training.

### Parameters setting

Table 1 exhibits the parameter settings for the experiment.

Two evaluation metrics, Bilingual Evaluation Understudy (BLEU) and Meteor, are used to comprehensively evaluate the performance of the FACT model[29–31]. These two metrics are among the most commonly used and representative automated evaluation tools in the current field of MT research. They have been widely applied in authoritative translation evaluation tasks such as the Workshop on Machine Translation (WMT), and have good universality and reliability. BLEU measures translation quality by calculating the n-gram match between the translated text and the reference answer. Specifically, BLEU calculates the precision of n-grams in the translated text, and its equation is as follows:

$$P_n = \frac{c_n}{r_n} \tag{18}$$

$P_n$ refers to the n-gram precision; $c_n$ represents the number of times the n-gram units in the translation match those in the reference answer; $r_n$ denotes the total number of n-gram units in the translation. The final BLEU score of the translation is the weighted average of the precision for each n-gram unit, which can be written as:

$$BLEU = \exp\left(\sum_{n=1}^{N} \omega_n \log P_n\right) \tag{19}$$

| Parameter category | Parameter | Value |
|---|---|---|
| Model structure | The number of encoder layers | 6 |
| | The number of bidirectional decoder layers | 5 |
| | Multimodal decoder | 1 |
| Model parameters | Word embedding dimension | 256 |
| | Hidden unit dimension | 512 |
| | Feedforward layer dimension | 2048 |
| | Optimizer | Adam |
| | Initial learning rate | 0.0001 |
| | Number of epochs | 50 |
| | Mini-batch size | 64 |
| | Warm-up | 4000 |
| | Dropout | 0.1 |

**Table 1**. Experimental parameter settings.

$\omega_n$ is the weighting factor for each n-gram unit. To avoid giving overly high scores to shorter translations, BLEU introduces a brevity penalty (BP) to adjust the score. The calculation of BP reads:

$$BP = \begin{cases} 1, & if\ r \geq c \\ \exp(1 - \frac{r}{c}), & if\ r < c \end{cases} \tag{20}$$

*r* and *c* represent the length of the reference and candidate translations. The final BLEU score is obtained by combining the BP of short sentences with the weighted average of n-gram precision, as follows:

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} \omega_n \log P_n) \tag{21}$$

The advantages of BLEU lie in its simplicity and speed of computation, making it suitable for large-scale evaluations. However, it relies solely on lexical-level matching, neglecting linguistic features such as semantic similarity and syntactic variations. As a result, it demonstrates limited effectiveness when handling synonyms, word order changes, or translations that maintain semantic consistency but are expressed differently.

In contrast to BLEU, Meteor adopts a word alignment-based evaluation method, which better considers semantic information and word order. Meteor establishes a one-to-one correspondence between the words in the candidate translation and the reference translation to calculate precision and recall. The expression is as follows:

$$P = \frac{m_w}{M_{hypothesis}} \tag{22}$$

$$R = \frac{m_w}{N_{reference}} \tag{23}$$

*P* represents the proportion of words in the translation that match the reference words; $m_w$ denotes the number of matched words; $M_{hypothesis}$ and $N_{reference}$ refer to the total number of words in the translation and the reference. *R* implies the proportion of words in the reference that match the words in the translation. Meteor calculates an F1 score by combining precision and recall, and gives higher weight to recall. The equation is as follows:

$$F_\beta = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \tag{24}$$

$\beta$ controls the weight between precision and recall. To better handle word order issues, Meteor also introduces a chunking mechanism that penalizes translations with word order mismatches, as given in Eq. (25):

$$Penalty = \frac{C_{hypothesis}}{C_{reference}} \tag{25}$$

$C_{hypothesis}$ and $C_{reference}$ represent the number of chunks in the translated text and the reference answer, respectively. The final Meteor score combines the F1 score with the word order penalty, and is calculated using Eq. (26):

$$Meteor\ Score = F_\beta - Penalty \tag{26}$$

Compared to BLEU, Meteor places greater emphasis on translation fluency, semantic retention, and linguistic naturalness, thus generally exhibiting higher correlation with human evaluation in simulations. By employing both BLEU and Meteor metrics simultaneously, a comprehensive evaluation of the FACT model's translation performance can be conducted from two dimensions: formal accuracy and semantic acceptability. This makes a more authentic reflection of its practical effectiveness in MMT.

## Performance evaluation

(1) Comparison of model performance

Five representative baseline models are selected for comparison to comprehensively evaluate the performance of the proposed FACT model in MNMT tasks. These models are Transformer, Latent Multimodal Machine Translation (LMMT), Dynamic Context-Driven Capsule Network for Multimodal Machine Translation (DMMT), Target-modulated Multimodal Machine Translation (TMMT), and Imagined Representation for Multimodal Machine Translation (IMMT). Among them, the Transformer model is a classic architecture in MT and, as a pure text baseline model, effectively verifies the performance gains brought by multimodal mechanisms. LMMT uses latent variables to model multimodal interactions, emphasizing the potential semantic expressive power of image-text fusion in the latent space. DMMT introduces a dynamic context capsule mechanism to enhance semantic coupling between modalities during translation. TMMT guides visual information to participate in the translation generation process under a target modulation mechanism, improving target alignment between modalities. IMMT attempts to use an "imagination" mechanism to generate intermediate

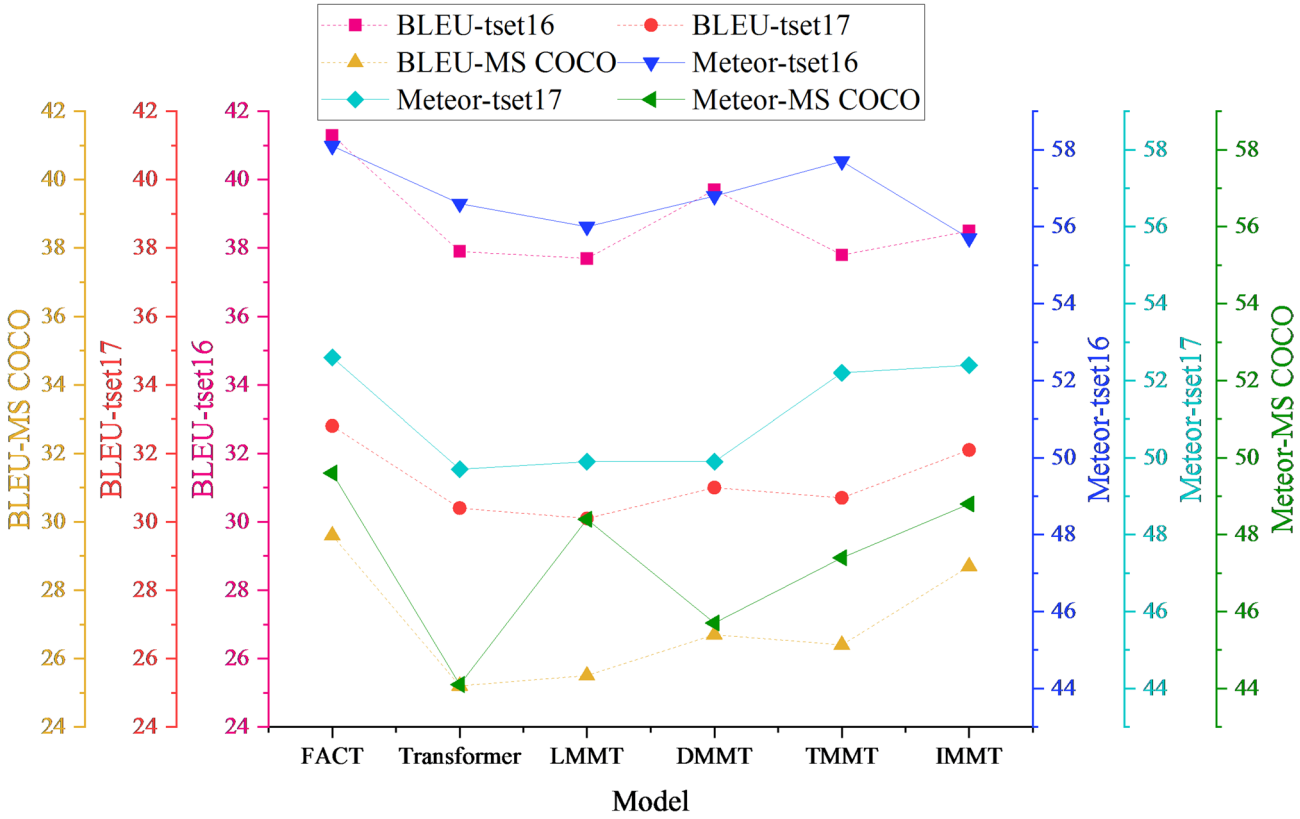


**Fig. 3**. Comparison of different models on the En-De translation task.

| Comparison model | *p* value |
|---|---|
| FACT and Transformer | 0.0023 |
| FACT and LMMT | 0.0013 |
| FACT and DMMT | 0.0020 |
| FACT and TMMT | 0.015 |
| FACT and IMMT | 0.028 |

**Table 2**. Significance test.

image representations for assisting semantic understanding and translation generation. All of the above models are representative methods in recent MNMT research, with strong representativeness and comparability. The primary reasons for not including large multimodal language models such as Generative Pre-trained Transformer 4 omni (GPT-4o) or Large Language and Vision Assistant (LLaVA) in this experiment are as follows. (1) These models are closed-source or commercialized, making fair comparisons under unified datasets and parameter configurations difficult; (2) Their training data and computing resources far exceed those accessible to the FACT model, rendering direct comparability infeasible; (3) FACT prioritizes structural lightness, training efficiency, and language learning adaptability over scale advantages. The above publicly structured and representative multimodal translation models are selected for horizontal comparison to ensure fair comparisons under unified datasets and parameter configurations. This enables more objective validation of the FACT model's performance advantages in semantic consistency modeling and future context information guidance. The BLEU and Meteor evaluation results of each model on the En-De translation task are depicted in Fig. 3. To further verify the statistical reliability of this advantage, a paired significance test is conducted on the performance scores between FACT and each benchmark model. The results are outlined in Table 2.

In Fig. 3, the FACT model proposed in this work outperforms other comparative models in both BLEU and Meteor scores. In the En-De translation tasks on the test16, test17, and MS COCO datasets, the FACT model achieves BLEU scores of 41.3, 32.8, and 29.6, respectively, which are significantly higher than those of the baseline models. In terms of Meteor scores, the FACT model also performs excellently, reaching 58.1, 52.6, and 49.6, outperforming other models. Although the performance of each model varies across different datasets, the FACT model consistently maintains leadership in BLEU and Meteor metrics, demonstrating its advantages in multimodal machine translation. Combined with Table 2, the *p* values of FACT compared with

| Model variant name | Model explanation |
|---|---|
| FACT-$L_{fd}$ | Removes the function $L_{fd}$ used to supervise the model's future target context information |
| FACT-$L_{md}$ | Removes the multimodal consistency loss function $L_{md}$ |
| FACT-$L_{fd}$-$L_{md}$ | Removes both $L_{fd}$ and $L_{md}$ |

**Table 3**. Names and descriptions of model variants.

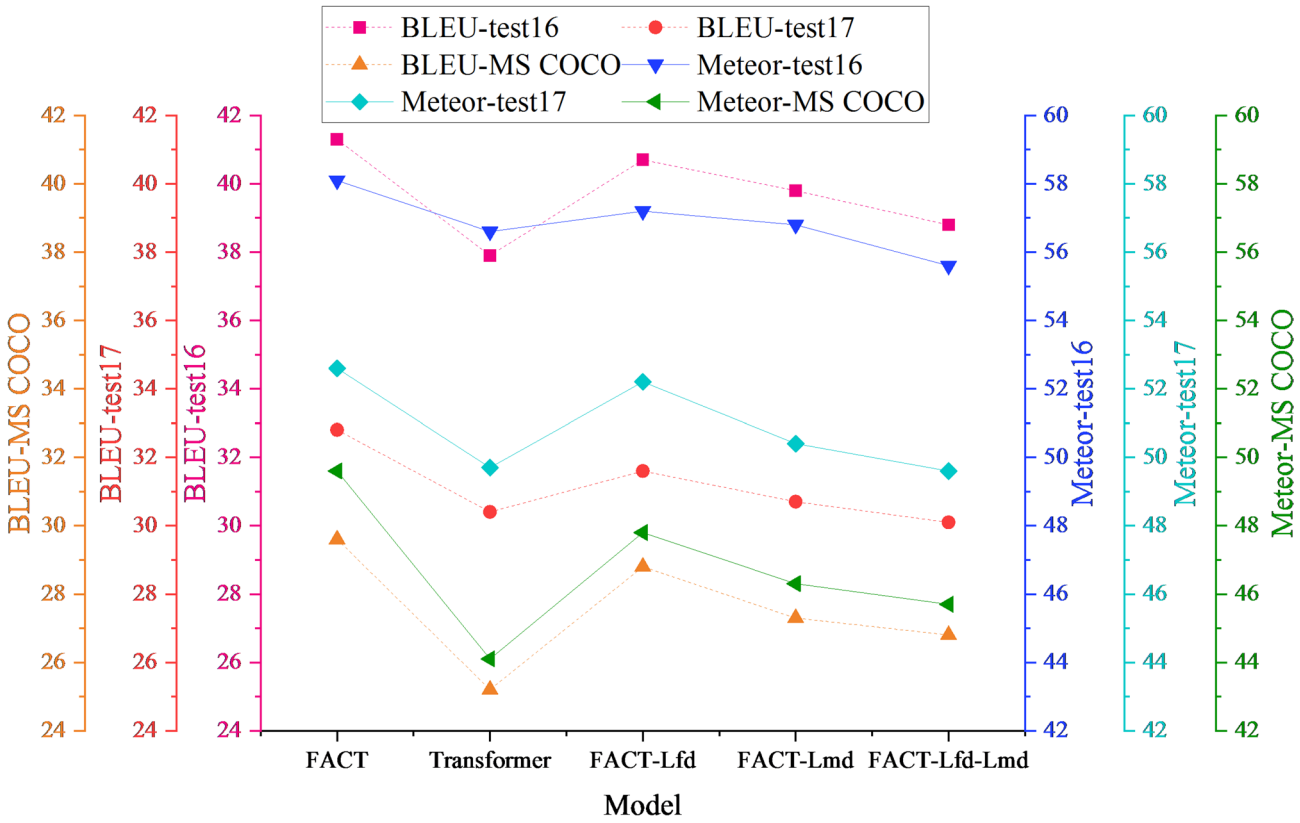


**Fig. 4**. Ablation experiment results on En-De translation task.

Transformer, LMMT, and DMMT are all less than 0.005, indicating highly significant performance differences. The $p$ values with TMMT and IMMT are 0.015 and 0.028, respectively, below the conventional significance level of 0.05. This demonstrates that FACT's performance advantages are statistically significant. The statistical results reveal that FACT remarkably outperforms all comparative methods in overall translation performance, fully confirming its effectiveness and advancement in MNMT. This is because the FACT model introduces two key innovations in structural design and modeling strategies compared to baseline models. On one hand, in future context information modeling, FACT leverages an attention-based future information guidance module to explicitly model the interaction among future target-side words, current source language, and visual features. Thus, it optimizes the directionality and contextual coherence of translation generation, which has not been systematically addressed in existing models. On the other hand, in multimodal consistency mechanisms, FACT constructs a loss function, strengthening the collaborative expressive capability between visual and linguistic modalities. By aligning the semantic space projections of images and texts, the collaborative expression ability between visual and language modalities is strengthened, and the robustness and generalization of image-text semantic fusion are improved. These two mechanisms complement each other, enabling FACT to outperform existing models in the granularity of information modeling and the depth of semantic alignment, significantly leading in multiple evaluation metrics such as BLEU and Meteor.

(2) Ablation experiment

Ablation experiments are conducted by creating the FACT model's variants to explore how this model integrates visual features to enhance translation performance. Table 3 lists the model variants.

Figure 4 demonstrates the results of ablation experiments on the En-De translation task, including BLEU and Meteor scores for the FACT model, three variant models, and the Transformer model. The "Transformer" in Fig. 4 is a pure text model without any image information or consistency modeling, serving as a baseline control.

Figure 4 reveals that for the En-De translation task, the BLEU and Meteor scores of the FACT model decrease when either the future target context information supervision function $L_{fd}$, or the multimodal consistency loss function $L_{md}$ is decreased. When both $L_{fd}$ and $L_{md}$ are removed, the FACT model's performance experiences the largest drop, but it still outperforms the Transformer model. Specifically, the BLEU scores decline by 6.05%, 8.23%, and 9.46% on the test16, test17, and MS COCO datasets. The Meteor scores decrease by 4.3%, 5.7%, and 7.86%, respectively. These results indicate that the future target context information and the multimodal consistency loss function remarkably influence the FACT model's translation performance.

Ablation experiments are also performed on the En-Fr and En-Cs translation tasks to verify the FACT model's generalization ability. Figures 5 and 6 show the results.

The results of the En-Fr translation task exhibit a similar pattern to the En-De findings. Both the future target context information supervision function $L_{fd}$ and the multimodal consistency loss function $L_{md}$ are deactivated. In this case, the FACT model achieves BLEU scores of 60.1, 53.0, and 43.8, and Meteor scores of 74.8, 70.1, and 63.7 on the test16, test17, and MS COCO datasets, respectively. These scores are all higher than those of the Transformer model.

Figure 6 shows that the results of the En-Cs translation task on the test2016 dataset are consistent with those of the En-De and En-Fr translation tasks. When the future target context information supervision function $L_{fd}$ and the multimodal consistency loss function $L_{md}$ are removed, the FACT model achieves BLEU and Meteor scores of 31.7 and 51.8, both exceeding those of the Transformer model. The results from En-Fr and En-Cs translation tasks further confirm that the FACT model can leverage multimodal consistency to learn future target context information, thus enhancing the performance of MMT.

(3) Impact of sentence length on model performance

The generated sentence lengths and BLEU scores for the FACT and Transformer models on the En-De translation task across the test16 and test17 datasets under varying source language sentence lengths are compared. Figure 7 presents the results.

Figure 7 shows that as the length of the source language sentence increases, the FACT model demonstrates a significant advantage in translation quality compared to the Transformer model. In the En-De translation task, the FACT model achieves a BLEU score of 44.1 for short sentences (0–10 words), outperforming the Transformer's 41.0. The translated sentence length is relatively short, with FACT producing a length of 8.4 and Transformer 8.2. As the source sentence length grows, the FACT model's translation quality advantage becomes even more pronounced. Additionally, its generated translation lengths adapt to the increase in source sentence length, producing more reasonable translation lengths for longer sentences. This indicates the model's strong

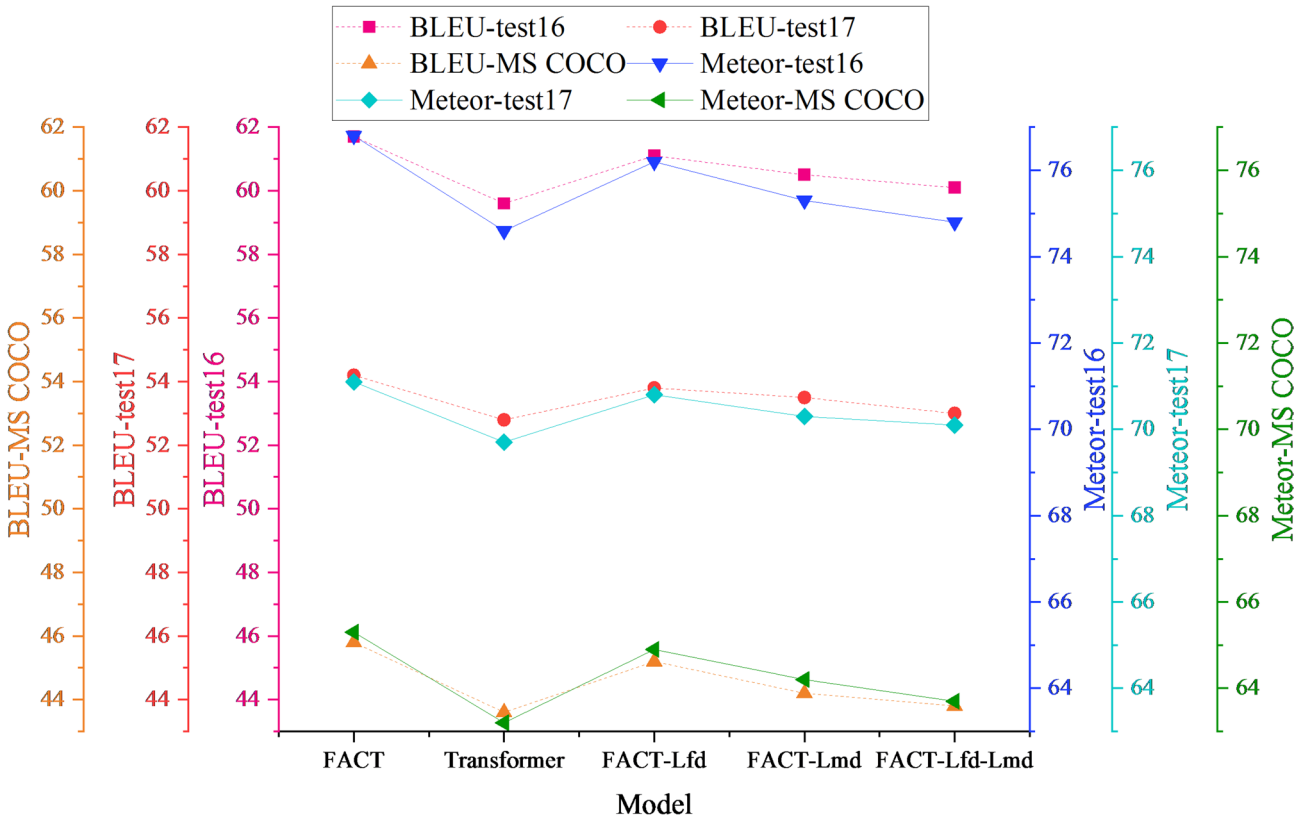


**Fig. 5**. Ablation experiment results on En-Fr translation task.

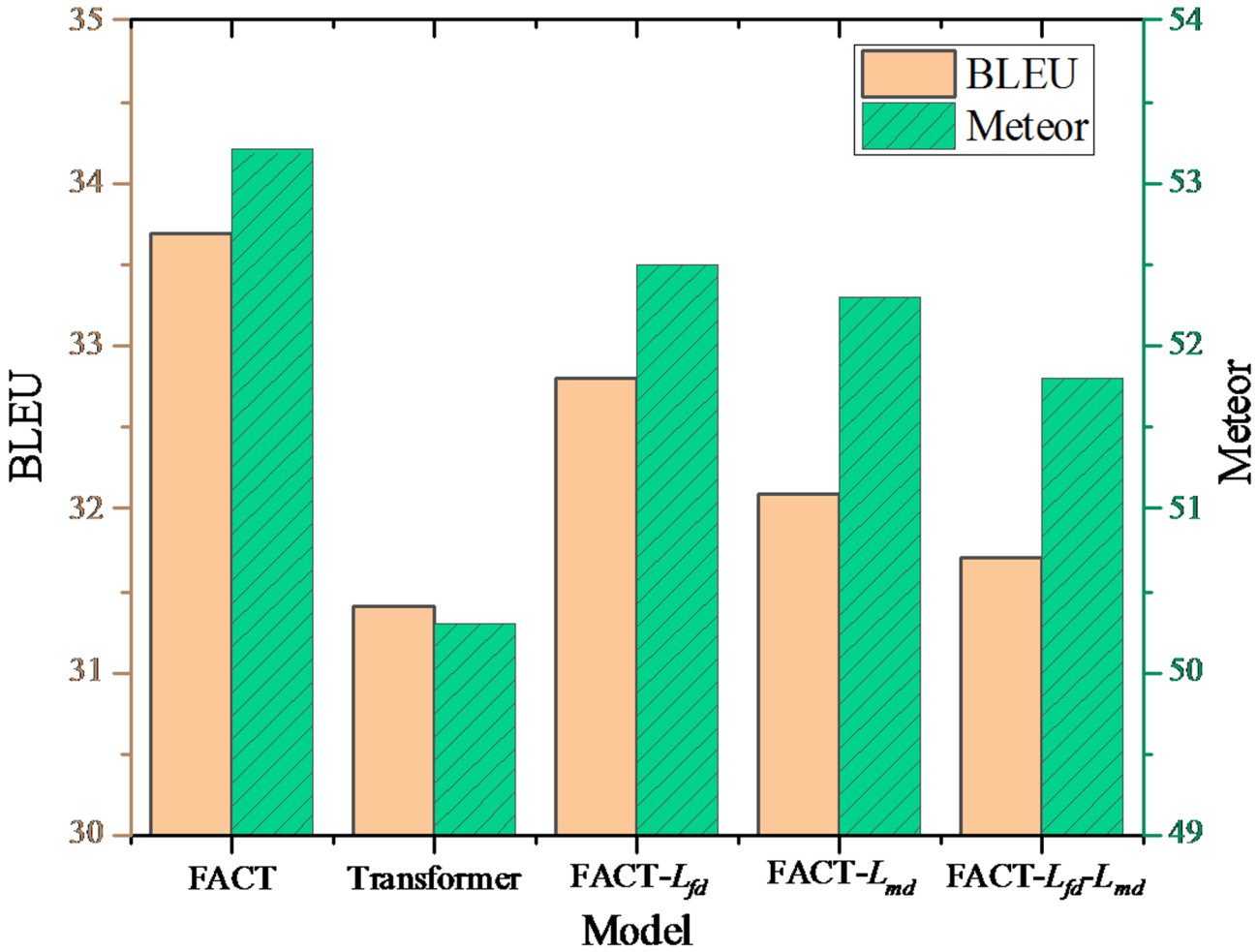


**Fig. 6**. Ablation experiment results on En-Cs translation task.

handling of long sentences. These findings demonstrate that the FACT model can more effectively predict future context when handling long sentence translation tasks, thereby improving translation quality.

(4) Impact of Model on Learning Investment Effect

To explore the effectiveness of the FACT model, experiments are conducted to evaluate its application in language learning. Figure 8 compares the learning process quality, learning efficiency, and learning outcomes between FACT and Transformer models.

Figure 8 suggests that the FACT model exhibits a distinct advantage over the Transformer model in language learning tasks. Specifically, it outperforms Transformer across multiple metrics, including learning efficiency, translation quality, user satisfaction, and understanding improvement. The learning efficiency of FACT is 83.2 words per hour, compared to 74.6 words per hour for the Transformer, highlighting FACT's potential to accelerate the learning process. Additionally, FACT achieves a translation quality score of 82.7, higher than the Transformer's 78.9, indicating its superior performance in translation quality. It also scores higher in both user satisfaction and understanding improvement. Overall, the FACT model offers higher efficiency and better learning outcomes in language learning tasks, demonstrating significant application potential.

## Discussion

The proposed FACT model demonstrates significant translation performance in MMT tasks. Particularly in the accuracy of translating long sentences, FACT exhibits greater robustness compared to other baseline models, making it more effective in handling complex translation tasks. Compared to previous research, several studies have achieved similar outcomes. For instance, Zhao et al. (2021) developed a word-region alignment-based MNMT model that improved BLEU scores by 1.0 on the Multi30k test set, enhancing text-visual semantic correlation[32]. Li et al. (2022) introduced an unsupervised multimodal translation model that analyzed object interactions in video through spatiotemporal graphs. Despite the absence of video data, this model still achieved good translation performance at both the sentence and word levels[33]. Tayir and Li (2024) designed an unsupervised multimodal translation model for low-resource language pairs. The model improved BLEU scores

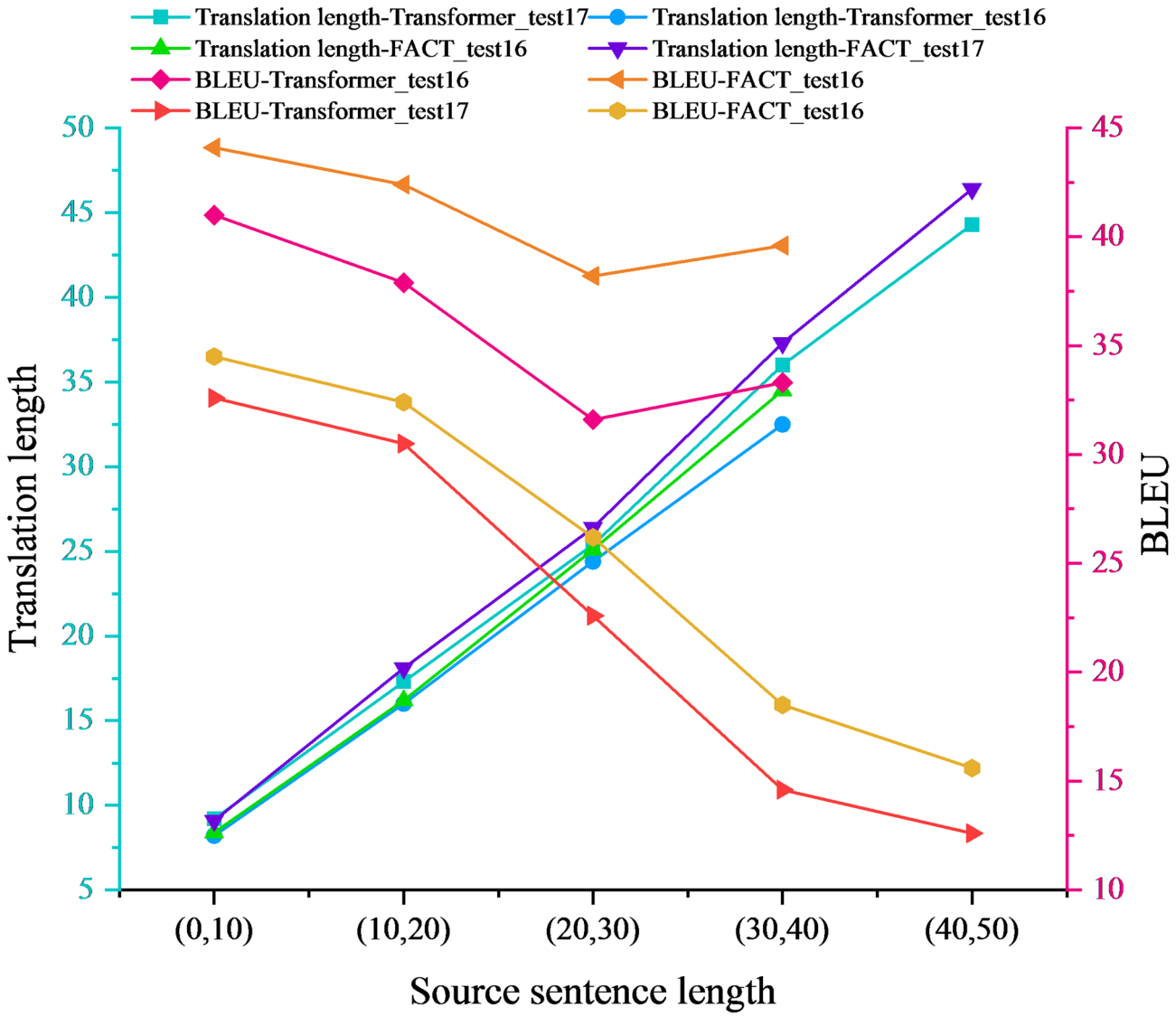


**Fig. 7**. Performance comparison of models at different source sentence lengths.

significantly in distant language pairs such as English-Uyghur and Chinese-Uyghur through transliteration, word-order reordering, and visual content masking[11]. These studies have demonstrated the potential of multimodal translation models across different datasets and tasks. The FACT model further optimizes the fusion of multimodal information by introducing future target context loss and multimodal consistency loss functions, resulting in higher performance in more challenging translation tasks. Compared to existing methods, the FACT model exhibits superior translation quality and higher user satisfaction, especially in handling complex translation tasks and long sentence translations.

More importantly, through the model design and optimization process of this study, several technically valuable insights have been gained. First, the modeling mechanism introducing future information indicates that the decoder does not have to entirely rely on autoregressive methods to construct the target sequence. Instead, it can indirectly model the target future context through guiding structures on the basis of maintaining training-inference consistency, thus enhancing the coherence and rationality of language generation. This mechanism provides new design ideas for future non-autoregressive translation or bidirectional generative translation models. Second, the multimodal consistency modeling approach focuses on semantic alignment between image and text latent spaces. This approach substantially reduces visual preprocessing requirements by avoiding explicit dependence on regional annotations or complex detectors, thereby enhancing model scalability and deployment practicality. These optimization strategies not only enhance translation performance but also balance model lightness and training efficiency. Additionally, the design concept of the FACT model exhibits strong transferability and generality. The future context modeling and modality consistency loss mechanism demonstrate broad applicability beyond image-text machine translation tasks. These approaches show significant potential for multimodal text generation and language teaching auxiliary systems. They are particularly suitable for high-precision semantic tasks, including medical image report generation and court document

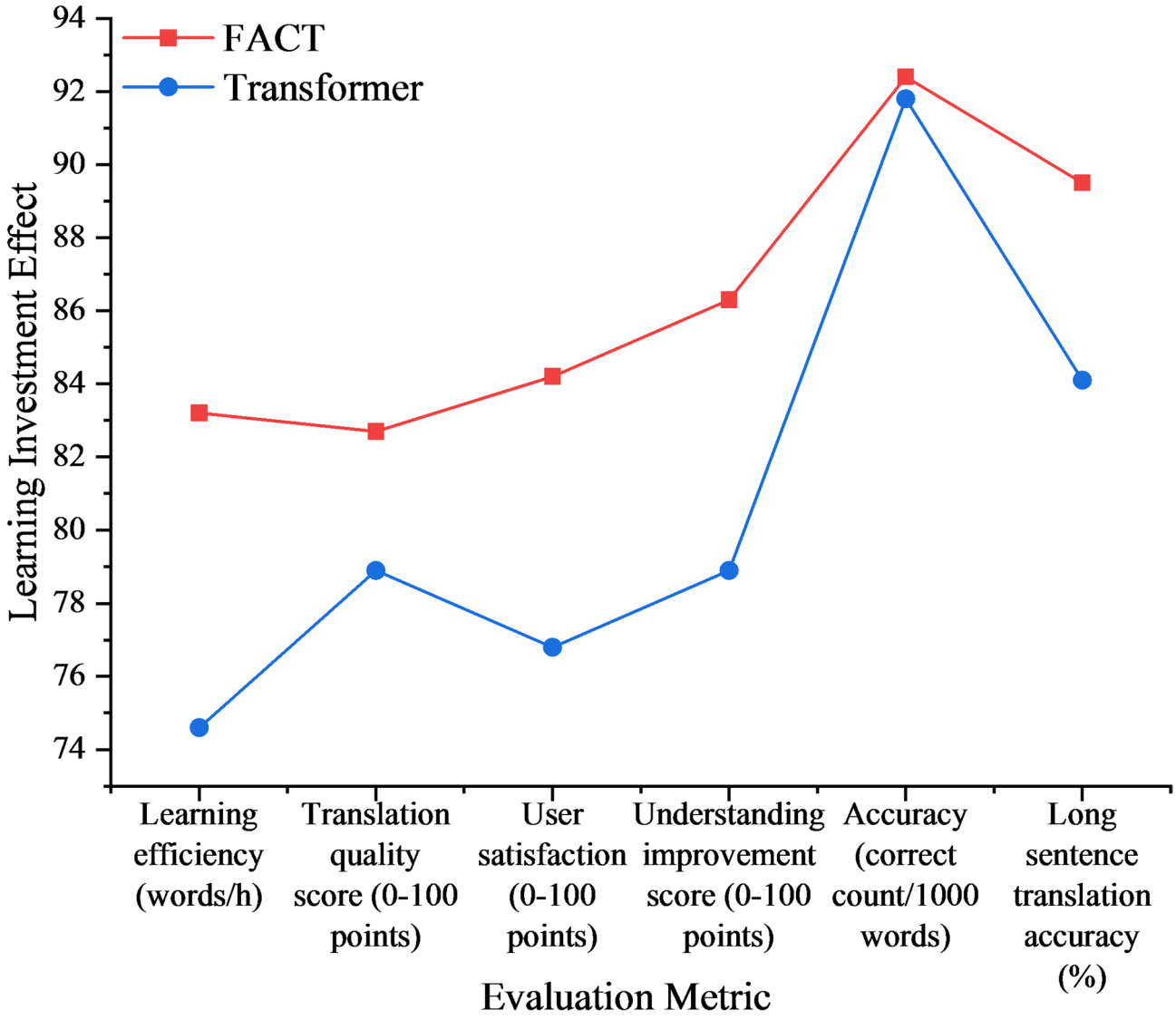


**Fig. 8**. Comparison of model impact on learning investment effect.

summarization. Especially in educational scenarios, the FACT model's enhancement of learning engagement effects indicates its promising application prospects in AI-assisted language learning platforms. Overall, the proposed multimodal modeling strategies and structural innovations improve the current translation systems' performance while providing referable theoretical foundations and engineering experience for cross-task migration of multimodal generation technologies.

## Conclusion

### Research contribution

This work proposes an MNMT model based on future information guidance and multimodal consistency modeling. The model's performance advantages are verified through experiments across various translation tasks, and its application effectiveness in language learning is also examined. The key contributions are as follows:

(1) In translation tasks, the FACT model outperforms other baseline models in BLEU and Meteor scores. On the tset16, tset17, and MS COCO datasets, FACT achieves BLEU scores of 41.3, 32.8, and 29.6, and Meteor scores of 58.1, 52.6, and 49.6, respectively, demonstrating its superiority in MMT. Meanwhile, the significance analysis of the performance scores between FACT and each baseline model shows that the $p$ values of the performance results comparison between FACT and other models are all 0.05, lower than the significance level. This indicates that the performance advantage of FACT is statistically significant. (2) Ablation experiments illustrate that in the FACT model, the future target context supervision function $L_{fd}$ and the multimodal consistency loss function $L_{md}$ are crucial for the model's translation performance. (3) In language learning tasks, the FACT model outperforms the Transformer model in learning efficiency, translation quality, user satisfaction, and understanding improvement scores. FACT's learning efficiency reaches 83.2 words/hour, higher

than Transformer's 8.6 words/hour. Metrics like translation quality and user satisfaction also show substantial advancements, highlighting their application potential during language learning.

### Future works and research limitations

The FACT model holds significant potential in MMT, effectively enhancing translation quality and learning efficiency. It also offers new insights for future research and the application of multimodal translation models. The limitations of this work mainly lie in the model's adaptability and generalization ability across various language pairs, which still require further validation, especially in low-resource language pairs. Future research could explore ways to improve the efficiency of multimodal information fusion, optimize the model's computational cost, and expand its application to more language pairs and domains.

## Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author Shuangshuang Lyu on reasonable request via e-mail zhangyan1219@jlenu.edu.cn.

Received: 9 January 2025; Accepted: 9 July 2025
Published online: 19 July 2025

## References

1. De Coster, M. et al. Machine translation from signed to spoken languages: State of the art and challenges. *Univ. Access Inf. Soc.* **23**(3), 1305–1331 (2024).
2. Liu, X. et al. A scenario-generic neural machine translation data augmentation method. *Electronics* **12**(10), 2320 (2023).
3. Mondal, S. K. et al. Machine translation and its evaluation: A study. *Artif. Intell. Rev.* **56**(9), 10137–10226 (2023).
4. Hu, J. Neural machine translation (NMT): Deep learning approaches through neural network models. *Appl. Comput. Eng.* **82**, 93–99 (2024).
5. NLLB Team. Scaling neural machine translation to 200 languages. *Nature* **630**(8018), 841 (2024).
6. Tonja, A. L. et al. Low-resource neural machine translation improvement using source-side monolingual data. *Appl. Sci.* **13**(2), 1201 (2023).
7. Almusharraf, A. & Bailey, D. Machine translation in language acquisition: A study on EFL students' perceptions and practices in Saudi Arabia and South Korea. *J. Comput. Assist. Learn.* **39**(6), 1988–2003 (2023).
8. Lee, S. M. The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Comput. Assist. Lang. Learn.* **36**(1–2), 103–125 (2023).
9. Wang, Y. Artificial Intelligence technologies in college English translation teaching. *J. Psycholinguist. Res.* **52**(5), 1525–1544 (2023).
10. Nam, W. & Jang, B. A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Syst. Appl.* **235**, 121168 (2024).
11. Tayir, T. & Li, L. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Trans. Asian Low-Resource Lang. Inf. Process.* **23**(4), 1–22 (2024).
12. Müller, C. & Rossi, A. Enhancing translation quality: A study on the impact of attention mechanisms in machine translation. *Eastern Eur. J. Multidiscipl. Res.* **3**(2), 47–52 (2024).
13. Safder, I. et al. Transforming language translation: a deep learning approach to Urdu-English translation. *J. Ambient. Intell. Humaniz. Comput.* **15**(10), 3651–3662 (2024).
14. Zhang, J. Research on intelligent translation system of spoken english based on cyclic neural network model. *Int. J. Inf. Commun. Technol. Educ. (IJICTE)* **20**(1), 1–16 (2024).
15. Klimova, B. et al. Neural machine translation in foreign language teaching and learning: a systematic review. *Educ. Inf. Technol.* **28**(1), 663–682 (2023).
16. Ohashi, L. Machine translation and language learning: teachers' perspectives and practices. In *Translation, Translanguaging and Machine Translation in Foreign Language Education*, 351–369 (2025).
17. Zhang, C. et al. Exploration of deep learning-driven multimodal information fusion frameworks and their application in lower limb motion recognition. *Biomed. Signal Process. Control* **96**, 106551 (2024).
18. Tomar, P. S., Mathur, K. & Suman, U. Fusing facial and speech cues for enhanced multimodal emotion recognition. *Int. J. Inf. Technol.* **16**(3), 1397–1405 (2024).
19. Guo, J., Su, R. & Ye, J. Multi-grained visual pivot-guided multi-modal neural machine translation with text-aware cross-modal contrastive disentangling. *Neural Netw.* **178**, 106403 (2024).
20. Li, L. et al. Multimodality information fusion for automated machine translation. *Inf. Fus.* **91**, 352–363 (2023).
21. Zhao, Y. et al. Region-attentive multimodal neural machine translation. *Neurocomputing* **476**, 1–13 (2022).
22. Yang, H. Optimized English Translation system using multi-level semantic extraction and text matching. *IEEE Access* (2024).
23. Guo, J. et al. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans. Inf. Syst. (TOIS)* **40**(4), 1–42 (2022).
24. Shi, X. & Yu, Z. Adding visual information to improve multimodal machine translation for low-resource language. *Math. Probl. Eng.* **2022**(1), 5483535 (2022).
25. Wang, H. et al. Progress in machine translation. *Engineering* **18**, 143–153 (2022).
26. Liu, C. & Xu, B. A night pavement crack detection method based on image-to-image translation. *Comput.-Aided Civil Infrastruct. Eng.* **37**(13), 1737–1753 (2022).
27. Hirasawa, T. et al. Pre-trained word embedding and language model improve multimodal machine translation: A case study in Multi30K. *IEEE Access* **10**, 67653–67668 (2022).
28. Tong, K. & Wu, Y. Rethinking PASCAL-VOC and MS-COCO dataset for small object detection. *J. Vis. Commun. Image Represent.* **93**, 103830 (2023).
29. Chauhan, S. et al. Adableu: A modified bleu score for morphologically rich languages. *IETE J. Res.* **69**(8), 5112–5123 (2023).
30. Han, C. & Lu, X. Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom?. *Comput. Assist. Lang. Learn.* **36**(5–6), 1064–1087 (2023).
31. Islam, M. A. & Mukta, M. S. H. A comprehensive understanding of popular machine translation evaluation metrics. *Int. J. Comput. Sci. Eng.* **25**(5), 467–478 (2022).
32. Zhao, Y. et al. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 244–259 (2021).
33. Li, M. et al. Video pivoting unsupervised multi-modal machine translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 3918–3932 (2022).

## Author contributions

Yan Zhang: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation. Shuangshuang Lyu: writing—review and editing, visualization, supervision, project administration, funding acquisition.

## Funding

This work was supported by the Research Project of Jilin Higher Education Association (No. JGJX24C130).

## Declarations

## Competing interests

The authors declare no competing interests.

## Ethics statement

This article does not contain any studies with human participants or animals performed by any of the authors. All methods were performed in accordance with relevant guidelines and regulations.

## Additional information

**Correspondence** and requests for materials should be addressed to S.L.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.