# Business English Translation Quality Assessment Using Multimodal BERT

**Shaojing Xiang**[1]
[1]Department of Languages and Humanities, Guangzhou College of Technology and Business, Guangzhou 510850, Guangdong, China
Corresponding author: Shaojing Xiang (e-mail: 13431569283@163.com)

**ABSTRACT** Business translation quality assessment faces numerous challenges, particularly for professional technical documents in engineering and industrial applications, where accurate terminology, numerical information, images, and structured tables are essential for reliable information transmission. Such multimodal consistency is also critical for documentation management and knowledge exchange in advanced electromagnetic systems and communication engineering. To address these issues, this paper proposes a multimodal BERT framework for no-reference translation quality assessment. The framework jointly integrates three modalities: text (source document and translation), images, and structured tables. Textual semantics are encoded using BERT, visual features are extracted by ResNet50, and tabular information is represented through TUTA. A multi-head cross-modal attention mechanism is then employed for lexical and cell alignment, while gated weighted pooling adaptively fuses heterogeneous modal representations to regress a unified translation quality score. To validate the proposed approach, a multimodal business dataset is constructed from publicly available reports issued by State Grid Corporation of China and Sinopec during 2023–2024, comprising approximately 12,000 documents, 48,000 sentence pairs, 20,000 image fragments, and 15,000 structured tables, with gold-standard annotations generated using an extended MQM evaluation scheme. Experimental results demonstrate that the proposed method achieves an RMSE of 0.06 for sentence-level quality regression, with Spearman and Pearson correlations of 0.87 and 0.88, respectively, relative to human assessment. In simulated business evaluation scenarios, the framework reduces economic risk by 38.64% and lowers the customer complaint rate to 3.21 %. These results demonstrate that multimodal BERT provides an effective and verifiable solution for translation quality assessment while offering valuable support for multilingual technical documentation, electromagnetic engineering communication, and intelligent information management in complex industrial environments.

**INDEX TERMS** multimodal BERT, translation quality assessment, cross-modal attention, electromagnetic engineering communication, intelligent information management

## I. INTRODUCTION

IN today's global business activities, cross-language communication and information exchange are becoming increasingly frequent. Business English translation plays a key role in scenarios such as contracts, invoices, reports, marketing materials, and technical documents. This challenge is profoundly felt in international trade-heavy sectors, where precise English translation is critical for manufacturing consistency, ensuring the accuracy of complex technical specifications and guaranteeing compliance with diverse international labeling regulations [1,2]. Accurate translation is not only related to the accuracy of information transmission but also directly related to the correctness of legal, financial and business decisions [3]. Existing translation quality assessment systems are mostly based on text similarity or automatic scoring/ quality estimation methods that only use text modality [4,5]. When faced with business materials containing pictures, tables, invoices or other visual/structured information, they often fail to capture key contextual clues, resulting in missed term mistranslations, inconsistent values and units, and semantic mismatches in tables. Business English has a high degree of terminology density, formal registers and strict requirements on values/formats. Therefore, the evaluator must not only determine semantic alignment, but also verify term consistency, value/unit fidelity and layout integrity. Multimodality refers to using text, images and tables (including layout/coordinate information) as complementary evidence sources to improve the detection ability of the above-mentioned business-specific errors through cross-modal alignment and fusion, and achieve more reliable reference-free quality estimation. Therefore, developing methods that can understand cross-modal context and improve the accuracy of business translation quality assessment has important theoretical significance and practical value. Existing research on translation quality assessment has gradually transitioned from rule-

based indicators to deep learning and large model-driven methods, and has made progress in error detection, context analysis and semantic similarity calculation. However, most research is still limited to a single text modality and lacks the comprehensive use of images, tables and layout information in business documents. There are deficiencies in dealing with term consistency, numerical verification and visual context maintenance in business English translation.

Multimodal methods are gradually being applied to translation and quality assessment research. Scholars are trying to integrate information such as text, images, and videos to improve translation performance and assessment accuracy. Han C proposed a unified framework of "multilingual, multimodal, and multi-agent", pointing out that research is still focused on text and English language pairs, and other modalities are underexplored [6]. Tayir T proposed an unsupervised multimodal method for low-resource and long-distance language pairs, applying visual content-assisted modeling, and significantly improved BLEU in English-Uyghur and Chinese-Uyghur translation [7]. Existing research has shown that multimodality has significant advantages in alleviating problems such as low resources, insufficient semantic alignment, and lack of contextual understanding. However, its application in specific fields such as business English is limited, and most of the methods are based on text and image modalities, and structured table information has not been included in the research.

This paper aims to build a deployable translation quality evaluator for real-world business scenarios. By jointly modeling evidence from three modalities—text, visual, and tabular—the model aims to improve the accuracy of discriminating and evaluating business English translations. It constructs and manually annotates a multimodal business dataset (approximately 48,000 sentence pairs, including image segments and table units) and uses MQM (Multidimensional Quality Metrics) for scoring. Technically, it uses BERT to extract textual semantics, a ResNet50 (Residual Network 50 layers) model to extract visual features, and a TUTA (Tree-based Transformer for Table Representation Learning) to encode the table structure. Linear projection, cross-modal attention, and gating fusion are then used to form a unified multimodal representation. Finally, it uses a regression head to concurrently optimize the weighted mean square error (WMSE), modality alignment, and ranking losses for end-to-end training and fine-tuning. This suppresses modalities affected by OCR or image noise and amplifies evidence directly relevant to the score. In large-scale comparative experiments, this method achieved an RMSE (Root Mean Square Error) of 0.06 and a MAE (Mean Absolute Error) of 0.04 in regression error, a 0.06 RMSE reduction compared to BLEURT (Bidirectional Encoder Representations from Transformers for Evaluation of Retrieval and Translation). Correlations with manual MQM reached Spearman's and Pearson's correlation coefficients of 0.87 and 0.88, respectively. In engineering evaluations, the model maintained an acceptable inference time (86.73 ms) and delivered significant business benefits (49.83% man-hour savings, 38.64% economic risk reduction, and 3.21% customer complaint rate), directly shortening delivery cycles, reducing labor and compliance costs, and improving customer satisfaction and corporate risk control capabilities. The results show that Multimodal BERT not only improves evaluation accuracy when processing business text containing charts and structured numbers but also has the potential for practical deployment and significant business value-added.

## II. MULTIMODAL BERT FRAMEWORK

### *TEXT SEMANTIC REPRESENTATION*

In a multimodal translation quality assessment system, text semantic representation constitutes the core input feature. This paper adopts BERT as a text representation model to encode the original business English text and its corresponding translation [8,9].

### *INPUT EMBEDDING CONSTRUCTION*

For a sequence of original business English text or translation, it is first mapped into subword units through the WordPiece tokenizer. Each subword is represented as the sum of three vectors, as shown in Equation (1).

$$x_j = e_j^{tok} + e_j^{seg} + e_j^{pos} \tag{1}$$

In formula (1), $e_j^{tok}$ represents the word embedding vector, $e_j^{seg}$ represents the sentence fragment embedding, which is used to distinguish the original text from the translation, and $e_j^{pos}$ represents the position embedding, which marks the word order.

### *CONTEXT DEPENDENCY MODELING*

The input sequence passes through BERT's multi-layer Transformer encoder, each layer of which includes a multi-head self-attention mechanism and a feedforward neural network (FFN) [10]. For the input matrix, the query, key, and value matrices are first calculated as shown in (2).

$$Q = H^{(l-1)}W_Q, \quad K = H^{(l-1)}W_K, \quad V = H^{(l-1)}W_V \tag{2}$$

In formula (2), $W_Q$, $W_K$, and $W_V$ represent the corresponding learnable parameters, and $H^{(l-1)}$ represents the input matrix.

The attention weight is calculated as shown in formula (3).

$$Attention(Q, K, V) = Softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

In formula (3), $\sqrt{d_k}$ represents the scaling factor.

The multi-head attention mechanism splits the attention into To heads, as shown in formula (4) [11].

$$\begin{cases} MultiHead(Q, K, V) = Concat(head_1, \ldots, head_{T_o})W_O, \\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \end{cases} \tag{4}$$

In formula (4), $W_O$ represents the weight matrix of the multi-head attention mechanism fusion. The output vector of each position is independently input into the feedforward network, as shown in formula (5).

$$FFN(h) = \max(0, hW_1 + b_1)W_2 + b_2 \tag{5}$$

In formula (5), $W_1$ and $W_2$ represent the weight matrices of the feedforward network layer, and $b_1$ and $b_2$ represent the corresponding bias terms. After $L$ layers of Transformer encoders, the context-enhanced representation $H^{(L)}$ is obtained.

### SEMANTIC VECTOR EXTRACTION

In BERT, the global semantic information of the sequence is provided by the final layer representation $h_{\text{cls}}^{(L)}$ of the special token [CLS] (Classification Token), which is denoted as $f$. To preserve word granularity information, this paper also uses average pooling to obtain the overall sentence vector, and the text representation vector $f_{\text{Text}}$ is denoted as $[f; f_{\text{avg}}]$.

In the business English translation quality assessment task, the original text and the translation are encoded separately, and the difference between the two is measured by vector similarity, combined with dot product similarity, as shown in formula (6). The final quality feature of the text modality is denoted as $F_{text} = [\Delta f; Sim]$.

$$Sim(Sen_{src}, Sen_{tgt}) = \frac{f_{src} \cdot f_{tgt}}{\|f_{src}\|\|f_{tgt}\|} \tag{6}$$

In formula (6), $Sen_{src}$ and $Sen_{tgt}$ represent the original sentence and the translated sentence respectively, and $f_{src}$ and $f_{tgt}$ represent the encoding results of the original and translated sentences, respectively.

### IMAGE FEATURE MODELING

In the task of evaluating the quality of business English translation, images often appear in the form of product displays, scanned copies of business contracts, or promotional advertisements, carrying semantic information that complements the text. To fully utilize this information in the evaluation process, this paper uses the ResNet50 model as a visual feature extractor and embeds the extracted multi-level image features into the multimodal BERT framework to achieve cross-modal semantic alignment.

In the ResNet50 model, the first convolution layer uses a 7×7 convolution kernel with a stride of 2 to extract low-level features from the normalized image, as shown in (7) [12,13].

$$\hat{X}^{(1)} = \sigma\left(\hat{I} * W^{(1)} + b^{(1)}\right) \tag{7}$$

In formula (7), $W^{(1)}$ represents the convolution kernel weight, $b^{(1)}$ represents the bias, $*$ represents the convolution operation, and $\sigma$ represents the ReLU activation function.

The core of the ResNet50 model is the residual block, whose calculation formula is shown in (8).

$$\begin{cases} B^{(l)} = \sigma\left(C\left(\hat{X}^{(l)}; W^{(l)}\right) + D\left(\hat{X}^{(l)}\right)\right), \\ C\left(\hat{X}^{(l)}; W^{(l)}\right) = W_3^{(l)} * \sigma\left(W_2^{(l)} * \sigma\left(W_1^{(l)} * \hat{X}^{(l)}\right)\right). \end{cases} \tag{8}$$

In formula (8), $C$ represents the residual mapping function consisting of three layers of convolution, $D$ represents the identity mapping, and $B^{(l)}$ represents the residual block output. $W_1^{(l)}$ represents the dimensionality reduction convolution, $W_2^{(l)}$ represents the spatial feature extraction, and $W_3^{(l)}$ represents the dimensionality increase convolution.

After 50 layers of residual calculation, the final high-dimensional feature map is obtained. Global average pooling is now applied to it, as shown in formula (9) [14,15].

$$p_k = \frac{1}{height_L \cdot width_L} \sum_{i=1}^{height_L} \sum_{j=1}^{width_L} \hat{X}_{i,j,k}^{(L)} \tag{9}$$

In formula (9), $\hat{X}_{i,j,k}^{(L)}$ represents the high-dimensional feature map. Finally, the image representation vector $P_{pic}$ of fixed dimension is obtained.

### STRUCTURED TABLE INFORMATION ENCODING

In business English translation scenarios, tables are often used to display contract terms, financial data, and product parameters. Their structured information is crucial for evaluating translation quality. This paper uses TUTA to model tables to capture their row and column structure, cell content, and cross-cell dependencies, generating a table semantic representation that can be uniformly aligned with text and images.

This paper first uses a word embedding matrix to encode cell text into a vector sequence, and then uses intracell self-attention to obtain a cell representation, as shown in (10).

$$u_{i,j} = Attn\left(\beta_1^{(i,j)}, \beta_2^{(i,j)}, \ldots, \beta_m^{(i,j)}\right) \tag{10}$$

In formula (10), $u_{i,j}$ represents the cell representation, and $\beta_m^{(i,j)}$ represents the vector after the cell text is encoded.

To model the row and column structure of the table, row encoding and column encoding are applied, respectively, as shown in formula (11).

$$o_i^{row} = W_r \cdot \gamma(i), \quad o_j^{col} = W_c \cdot \gamma(j) \tag{11}$$

In formula (11), $\gamma(i)$ and $\gamma(j)$ represent the one-hot row and column index vectors, respectively, and $W_r$ and $W_c$ represent the learnable weights.

The final input representation $RL_{i,j}^{(0)}$ of the cell is shown in formula (12).

$$RL_{i,j}^{(0)} = u_{i,j} + o_i^{row} + o_j^{col} \tag{12}$$

TUTA applies two parallel modules, row attention and column attention, into the table structure, as shown in (13).

$$\begin{cases} \alpha^{row}_{(i,j),(i,k)} = \dfrac{\exp\left(\frac{(RL^{(l)}_{i,j}W_q)\cdot(RL^{(l)}_{i,k}W^T_k)}{\sqrt{d_h}}\right)}{\sum^{Col}_{k'=1}\exp\left(\frac{(RL^{(l)}_{i,j}W_q)\cdot(RL^{(l)}_{i,k'}W^T_k)}{\sqrt{d_h}}\right)}, \\ \alpha^{col}_{(i,j),(k,j)} = \dfrac{\exp\left(\frac{(RL^{(l)}_{i,j}W_q)\cdot(RL^{(l)}_{k,j}W^T_k)}{\sqrt{d_h}}\right)}{\sum^{Row}_{k'=1}\exp\left(\frac{(RL)W_q\cdot(RL^{(l)}_{k',j}W^T_k)}{\sqrt{d_h}}\right)}. \end{cases} \tag{13}$$

In formula (13), $\alpha^{row}_{(i,j),(i,k)}$ and $\alpha^{col}_{(i,j),(k,j)}$ represent the attention weights within the row and column, respectively.

The updated cell representation is shown in formula (14).

$$\widetilde{RL}^{(l)}_{i,j} = \sum_{k=1}^{Col}\alpha^{row}_{(i,j),(i,k)}\cdot\left(RL^{(l)}_{i,k}W_V\right) + \sum_{k=1}^{Row}\alpha^{col}_{(i,j),(k,j)}\cdot\left(RL^{(l)}_{k,j}W_V\right) \tag{14}$$

In formula (14), $\widetilde{RL}^{(l)}_{i,j}$ represents the updated cell representation. To capture the global information of the table, TUTA further applies a table-level aggregation operator to calculate the global vector representation of the table, as shown in formula (15).

$$RQ = \frac{1}{Row\cdot Col}\sum_{i=1}^{Row}\sum_{j=1}^{Col}\widetilde{RL}^{(L)}_{i,j} \tag{15}$$

The final encoding result of the structured information table is the concatenation of the table's global vector representation and the cell-level representation, defined as $R_{\text{table}}$.

To further meet the stringent requirements for numerical accuracy (such as measurement tolerances, units of measurement, and monetary discrepancies) in commercial quality control reports, this paper introduces a numerical structure-sensitive enhancement mechanism based on TUTA encoding. For cells containing key numerical values, they are first split into structured numerical fragments using a regularized numerical tokenizer (including decimal points, thousands separators, negative signs, and unit symbols). An additional "numerical accuracy embedding" is then overlaid at the input layer to explicitly mark the dimensions, digit length, and format characteristics of the numerical values. Furthermore, TUTA incorporates "numerical consistency weights" in the row/column attention calculation stage, making the system more sensitive to attention shifts caused by numerical anomalies (such as digit anomalies, unit mismatches, and monetary deviations), thus highlighting potential errors in cross-cell comparisons. Simultaneously, to achieve cross-modal evidence association, this paper adds numerical alignment constraints to the cross-modal attention layer, enabling key numerical cells to correspond with monetary regions in images, numerical text blocks in table screenshots, and numerical entities in the original text. If the translation contains "1,234.5" which is inconsistent with the original image containing "1,234.56", the model can capture the source of error through numerical precision encoding and visual region alignment.

The multimodal BERT framework is shown in Figure 1.

Figure 1 illustrates the overall structure of the multimodal BERT framework. This framework first divides the input into three modalities: text, image, and structured table information. Feature extraction is performed using BERT, a ResNet50 model, and TUTA, respectively, to capture semantic, visual, and table contextual features. These three features are then deeply interacted and weighted in a multimodal fusion layer to form a unified multimodal representation. Finally, a quality assessment regression head maps the fused features and outputs a translation quality score, enabling comprehensive quality assessment of business English translations.

### *MULTIMODAL FEATURE FUSION AND TRANSLATION QUALITY SCORE REGRESSION HEAD DESIGN*

After completing the independent representation of text, images, and tables, a multimodal feature fusion module is constructed to capture cross-modal interaction information and output the translation quality score through the regression head. This paper adopts an attention-driven multimodal interaction layer + unified embedding space mapping + regression network approach to achieve this [16,17].

### *MODAL VECTOR NORMALIZATION AND PROJECTION*

To achieve modal unification, linear projection is used for processing, as shown in formula (16).

$$\begin{aligned} \widehat{F_{text}} &= F_{text}W_{F_{text}} + b_{F_{text}}, \\ \widehat{P_{pic}} &= P_{pic}W_{P_{pic}} + b_{P_{pic}}, \\ \widehat{R_{table}} &= R_{table}W_{R_{table}} + b_{R_{table}}. \end{aligned} \tag{16}$$

In formula (16), $W_{F_{text}}$, $W_{P_{pic}}$, and $W_{R_{table}}$ represent the weight matrices corresponding to text, image, and structured table, and $b_{F_{text}}$, $b_{P_{pic}}$, and $b_{R_{table}}$ represent the corresponding biases.

Before the cross-modal attention step, the table features from TUTA are first treated as a cell-level sequence. Each cell vector undergoes linear projection to uniformly adjust its dimensions to the same vector dimension as the BERT-encoded text features. Specifically, the cells in each row of the table are expanded from left to right and from top to bottom to form a fixed-length sequence; if the number of cells is less than the set maximum length, zero vectors are used to pad to ensure consistent sequence length. After dimensionality projection and serialization, the table vectors and text word vectors are aligned in the same embedding space, allowing them to directly participate in cross-modal attention calculations and achieve one-to-one or many-to-many interactions between text and table cells. This approach preserves the structured information of the table while ensuring the operability and stability of cross-modal feature fusion.

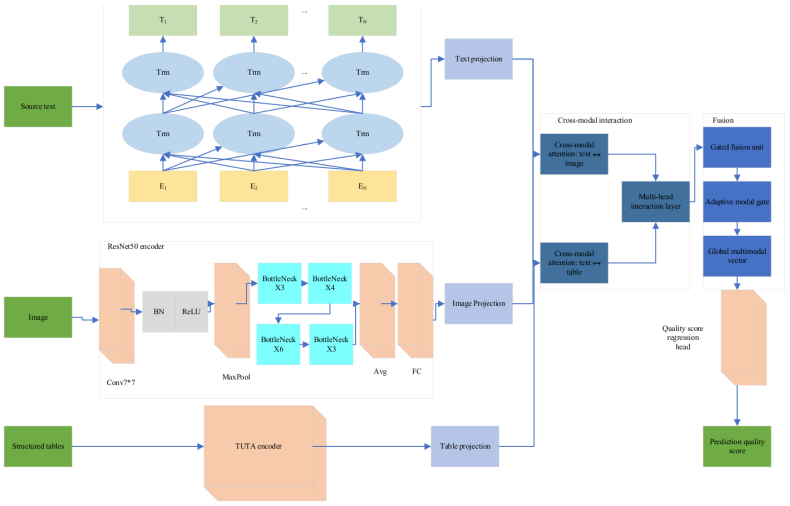


**FIGURE 1.** Multimodal BERT Framework

### *CROSS-MODAL ATTENTION MECHANISM*

To model inter-modal interactions, cross-modal attention is used, with text as the query and images and tables as the key values. The cross-modal attention weights are shown in (17) [18,19].

$$\begin{cases} \alpha_{\delta,\zeta}^{(F_{text}P_{pic})} = \dfrac{\exp\left(\dfrac{(\widehat{F_{text\delta}}W_Q^{Ftext})\cdot(\widehat{P_{pic\zeta}}W_K^{Ppic})^T}{\sqrt{d_h}}\right)}{\sum_{\zeta'=1}^{L_{Ppic}}\exp\left(\dfrac{(\widehat{F_{text\delta}}W_Q^{Ftext})\cdot(\widehat{P_{pic\zeta'}}W_K^{Ppic})^T}{\sqrt{d_h}}\right)}, \\ \alpha_{\delta,\vartheta}^{(F_{text}R_{table})} = \dfrac{\exp\left(\dfrac{(\widehat{F_{text\delta}}W_Q^{Ftext})\cdot(\widehat{R_{table\vartheta}}W_K^{Rtable})^T}{\sqrt{d_h}}\right)}{\sum_{\vartheta'=1}^{L_{Rtable}}\exp\left(\dfrac{(\widehat{F_{text\delta}}W_Q^{Ftext})\cdot(\widehat{R_{table\vartheta'}}W_K^{Rtable})^T}{\sqrt{d_h}}\right)}. \end{cases} \tag{17}$$

In formula (17), $\alpha_{\delta,\zeta}^{(F_{text}P_{pic})}$ and $\alpha_{\delta,\vartheta}^{(F_{text}R_{table})}$ represent the cross-modal attention weights for images and tables as keys, respectively, and $W_Q^{Ftext}$, $W_K^{Ppic}$, and $W_K^{Rtable}$ represent the query and key/value projection matrices.

The updated text representation is shown in formula (18).

$$\begin{aligned} \overline{F_{text\delta}} = &\sum_{\zeta=1}^{L_{Ppic}} \alpha_{\delta,\zeta}^{(F_{text}P_{pic})} \cdot (\widehat{P_{pic\zeta}}W_V) \\ &+ \sum_{\vartheta=1}^{L_{Rtable}} \alpha_{\delta,\vartheta}^{(F_{text}R_{table})} \cdot (\widehat{R_{table\vartheta}}W_V) \end{aligned} \tag{18}$$

In formula (18), $\overline{F_{text\delta}}$ represents the updated text representation. The cross-modal attention mechanism first treats each word in the text modality as a query, and each region of the image feature map or each cell of the table as a key and value. Attention calculation is applied directly between each word and each visual/table unit, forming a corresponding weight matrix of word-image region or word-table unit. These attention weights reflect the interaction strength between text and features from other modalities. In the multi-head mechanism, different heads learn different alignment patterns, such as local detail alignment, global semantic correspondence, or numerical precision alignment. Subsequently, the attention outputs of each head are concatenated and fused through a linear transformation to obtain the updated text representation. There are no intermediate aggregation steps in this process, ensuring that each text word can directly perceive the diverse features of the image or table unit, thereby accurately capturing key cross-modal semantic information and providing strong contextual support for subsequent translation quality assessment.

The cross-modal attention heatmap is shown in Figure 2.

Figure 2 shows a significant concentration of attention weights between text tokens and image regions. For example, "USD" aligns with the center of the image (R2C2), "Invoice" corresponds to R1C1, and "A123" corresponds to the lower-right region (R3C3), demonstrating that the model accurately captures the correspondence between semantic entities and visual evidence. Attention between text tokens and table cells exhibits a distinct structural pattern. For example, "Total" aligns with cell C4R3 in the last column of the table, and "1,234.56" corresponds to the specific numerical value in cell C4R3, demonstrating the model's ability to accurately map structured modalities. Overall, the cross-modal attention mechanism effectively enhances text's ability to perceive multimodal context, enabling the establishment of correspondences between key semantic information across modalities and providing solid support for subsequent fine-grained evaluation of translation quality.

### *MODALITY-LEVEL GLOBAL AGGREGATION*

In order to obtain the modal-level global vector, this paper performs weighted pooling on each modal feature, as shown in (19).

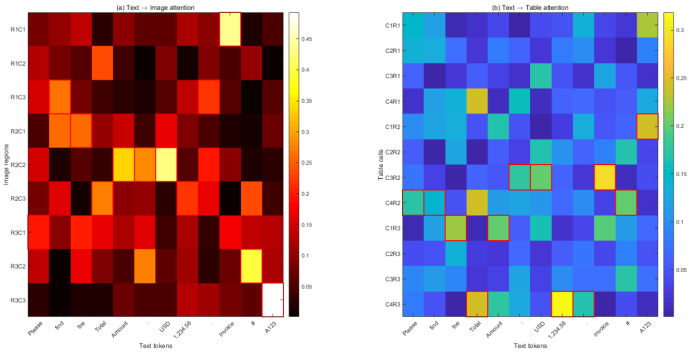

FIGURE 2. Cross-modal attention heatmap: (a) Text and image attention; (b) Text and table attention

$$
\begin{aligned}
z_{F_{text}} &= \frac{1}{L_{F_{text}}} \sum_{\delta=1}^{L_{F_{text}}} \overline{F_{text\delta}}, \\
z_{P_{pic}} &= \frac{1}{L_{P_{pic}}} \sum_{\zeta=1}^{L_{P_{pic}}} \widehat{P_{pic\zeta}}, \\
z_{R_{table}} &= \frac{1}{L_{R_{table}}} \sum_{\vartheta=1}^{L_{R_{table}}} \widehat{R_{table\vartheta}}.
\end{aligned} \tag{19}
$$

In formula (19), $z_{F_{text}}$, $z_{P_{pic}}$, and $z_{R_{table}}$ represent the weighted pooling features of each modality. Then a gating mechanism is used to achieve modal fusion, as shown in formula (20).

$$
\begin{cases}
\lambda_{F_{text}}, \lambda_{P_{pic}}, \lambda_{R_{table}} = Softmax\left([z_{F_{text}}, z_{P_{pic}}, z_{R_{table}}]W_\lambda + b_\lambda\right), \\
z = \lambda_{F_{text}} z_{F_{text}} + \lambda_{P_{pic}} z_{P_{pic}} + \lambda_{R_{table}} z_{R_{table}}.
\end{cases} \tag{20}
$$

In formula (20), $\lambda_{F_{text}}$, $\lambda_{P_{pic}}$, and $\lambda_{R_{table}}$ represent the weights corresponding to text, image, and structured table information, and $z$ represents the final fused features.

### TRANSLATION QUALITY SCORE REGRESSION HEAD DESIGN

The final fusion vector is input into the regression head to predict the translation quality score. The regression head consists of a two-layer fully connected network, as shown in (21).

$$
g = \sigma(zW_{out1} + b_{out1}), \quad \hat{y} = gW_{out2} + b_{out2} \tag{21}
$$

In formula (21), $W_{out1}$ and $W_{out2}$ represent the weights of the output layer, $b_{out1}$ and $b_{out2}$ represent the bias of the output layer, and $\hat{y}$ represents the predicted translation quality score output by the regression head.

### MODEL TRAINING AND OPTIMIZATION

In order to measure the difference between the prediction and the true score, the weighted mean square error is used as shown in (22).

$$
Loss_{wmse} = \frac{1}{N} \sum_{i=1}^{N} w_{ss} \cdot (y_i - \hat{y}_i)^2 \tag{22}
$$

In formula (22), $w_{ss}$ represents the sample weight and $y_i$ represents the true score of manual annotation. The modality alignment loss is shown in formula (23).

$$
Loss_{align} = -\sum_{(m_1,m_2)} \log \frac{\exp(\cos(z_{m_1}, z_{m_2})/\pi)}{\sum_{m' \neq m_1} \exp(\cos(z_{m_1}, z_{m'})/\pi)} \tag{23}
$$

In formula (23), $\pi$ represents the temperature parameter, cos represents the cosine similarity, and $m_1$ and $m_2$ belong to the three modes.

The final loss function is shown in formula (24).

$$
Loss = Loss_{wmse} + \eta_1 Loss_{align} + \eta_2 \|\theta\|_2^2 \tag{24}
$$

In formula (24), $\theta$ represents the set of all model parameters, $\|\theta\|_2^2$ represents the regularization term, and $\eta_1$ and $\eta_2$ represent the weight hyperparameters for modality alignment and regularization, respectively. To achieve efficient training, the AdamW optimizer is used to avoid overcontraction caused by gradient noise [20,21]. A Warmup + Cosine Annealing strategy is employed to prevent oscillation in the early stages of training and local minima in the later stages[22,23]. The learning rate is gradually increased in the early stages to ensure stable alignment of the multimodal feature space; it is gradually reduced in the later stages to prevent overfitting and improve the model's ability to capture detailed information.

The hyperparameter settings are shown in Table 1.

**TABLE 1.** Hyperparameters

| Hyperparameters | Value | Hyperparameters | Value |
|---|---|---|---|
| Number of training epochs | 10 | Number of attention heads | 8 |
| Batch size | 32 | Dropout ratio | 0.1 |
| Regression head hidden dimension | 512 | Gradient accumulation | 1 |
| Text learning rate | 2.00E-05 | Alignment loss weight | 0.1 |
| Visual/table learning rate | 1.00E-04 | L2 regularization coefficient | 1.00E-05 |
| Number of fusion layers | 4 | Weight decay | 0.01 |
| Maximum sequence length | 256 | Comparison temperature | 0.07 |
| Image size | 224 × 224 | Random seed | 42 |
| ResNet output dimension | 2048 | Gradient clipping threshold | 1 |
| TUTA unit representation dimension | 768 | Validation interval | 1000 step |
| Multimodal shared dimension | 768 | Early stopping rounds | No improvement after 5 verification |

## III. EXPERIMENTAL SYSTEM FOR BUSINESS TRANSLATION QUALITY ASSESSMENT

### *EXPERIMENTAL DATA AND PREPROCESSING*

This study constructs a multimodal translation quality assessment dataset for business scenarios. The data primarily comes from public reports published on the official websites of two medium-to-large multinational trading companies, State Grid Corporation of China and Sinopec, from 2023 to 2024. These include financial statements, product brochures, press releases, and sustainability reports. Translation candidates generated from the original English corporate documents using a commercial machine translation engine served as a low- and high-quality control sample. The data types include the source English text, the corresponding target language translation (primarily in Chinese), document scans (including screenshots of charts and graphs), and structured tables (either the original table in cell units or the table structure recovered from scans). The dataset consists of 12,000 business documents, containing approximately 48,000 sentence pairs (sourcetranslation pairs at the sentence level), approximately 20,000 relevant image fragments (charts, photos, and scanned page sections), and 15,000 sets of structured table cells (each table averaging 6–12 cells). All sentence pairs are manually quality-assessed using MQM-style hierarchical annotation (including error type labels: terminology, value/unit, format, register/fluency, etc.), and given a quality score of 0–100. Each sentence pair is independently scored by three reviewers with business translation experience, and a gold standard score is generated by combining majority vote and the mean. The total number of manually scored samples is 48,000. The data is divided into a training set of 38,400 sentences (80%), a validation set of 4,800 sentences (10%), and a test set of 4,800 sentences (10%).

The dataset constructed in this paper has a high coverage of both image and tabular modalities. Image data includes screenshots of contract pages, product display photos, flowcharts, and charts from sustainability reports, covering different resolutions, scan qualities, and layout styles. Tabular data comes from financial statements, product parameter tables, and contract clause tables, including single-row, multi-row/multicolumn merged cells, and structural layouts with varying column widths and row heights. Through this diverse data collection, the training set can cover common business document layouts and visual presentation formats, thereby enhancing the multimodal BERT model's generalization ability to various image and table types and ensuring robust handling of different document sources and formats in practical business translation quality assessments.

### *EXPERIMENTAL PLAN*

This study conducts experimental verification on the multimodal business dataset constructed in the previous paper (38,400 training sets, 4,800 validation sets, and 4,800 test sets), using manual MQM style scoring (0–100) as the gold standard. The comparison methods are divided into two categories: one is unimodal/text-based evaluation metrics and semantic scorers, including BERTScore (BERT), MoverScore (RoBERTa (Robustly Optimized BERT Pretraining Approach), BLEURT, and COMET (Crosslingual Optimized Metric for Evaluation of Translation); the other is multimodal or joint visual-text models, including VisualBERT, UNITER (Universal Image-TExt Representation), LXMERT (Learning Cross-Modality Encoder Representations from Transformers), and this paper's multimodal BERT (BERT+ResNet50+TUTA). BERTScore, MoverScore, BLEURT, and COMET are used as "metrics/scoring functions" to directly calculate scores for source-translation pairs and compare their correlation with manual scores. VisualBERT/UNITER/LXMERT and multimodal BERT are fine-tuned for regression (regression head predicts quality scores) on the same train-validation-test split, and training is performed using AdamW, Warmup + Cosine learning rate scheduling, mini-batch training with early stopping on the validation set are used. Hyperparameters are uniformly determined by grid search on the validation set to ensure comparability.

### *SIMULATED BUSINESS INDICATOR EVALUATION METHOD*

To verify the value of the proposed multimodal translation quality evaluation method in actual business scenarios, this paper constructs a simulated business indicator evaluation system to quantify the potential impact of the model on operational efficiency, risk control, and customer experience. This system includes four core indicators: time saving rate, process automation rate improvement, avoidable economic risk reduction rate, and customer complaint rate. Each indicator is evaluated based on a simulated environment of the company's real business processes, supplemented

by historical data verification to ensure the credibility and interpretability of the evaluation results.

(1) Simulation Environment Construction The simulation environment is based on the actual business processes of two medium-to-large multinational trading companies, including financial statement review, product manual translation, and cross-language release processes for press releases and sustainability reports. By decomposing the original processes, the execution time, manual operation steps, and potential error types for each task are defined. The processing time, error rate, and frequency of manual intervention for each task are initialized based on historical data and empirical values. During the simulation, the translation tasks are replaced by translated texts generated by different models instead of manual drafts. The system determines whether manual review or direct release is required based on the quality score predicted by the model, thus having different impacts on business efficiency and risk.

(2) Indicator Definition and Calculation Method

Time Saving Rate: Compares the percentage change in total time required for manual operation before and after model intervention. The more times manual review is reduced, the higher the time saved.

Process Automation Rate Improvement: Measures the increase in the proportion of steps in the process that do not require manual intervention due to the model. The higher the automation rate, the stronger the model's ability to replace manual labor in actual operation.

Avoidable Economic Risk Reduction Rate: Measures the extent of risk reduction by simulating potential errors and losses caused by translation quality. Risk assessment includes potential economic losses from terminology errors, amount/unit errors, and mistranslation of important clauses. The higher the model's prediction accuracy, the greater the probability of errors being detected or avoided in advance, thereby reducing potential losses.

Customer Complaint Rate: Simulates customer feedback on the readability, professionalism, and information completeness of the translated text, and calculates the complaint rate based on the model score and the proportion of manual review. The closer the model score is to the human gold standard, the lower the complaint rate.

(3) Risk Calculation Method The quantification of avoidable economic risk is based on the following principles:

Error Exposure Mechanism: In simulated business operations, translation errors may lead to contract misunderstandings, pricing errors, or inaccurate financial data. An error loss mapping table is constructed by statistically analyzing historical error frequencies and related economic impacts.

Model intervention effect: High-quality model predictions can directly replace manual review, reducing the probability of error exposure. Low-quality predictions still require manual intervention, and the error rate is retained based on historical data.

Cumulative risk reduction rate: Throughout the simulation process, the potential losses reduced by the model are compared with the total potential losses to obtain the percentage of avoidable economic risks. This calculation is iterated multiple times in the simulation environment to reduce the impact of individual task differences and ensure stable and reliable results.

(4) Transparency and Reproducibility To ensure the transparency of the evaluation, each simulation step, manual operation assumption, task execution time, error probability, and economic impact are recorded in the simulation system and can be exported as an analysis report. By comparing the changes in indicators of different models under the same simulation environment, the commercial value of each method can be quantified fairly and repeatably.

## IV. ANALYSIS OF TRANSLATION QUALITY ASSESSMENT RESULTS

### *BUSINESS ENGLISH TRANSLATION QUALITY SCORE REGRESSION PERFORMANCE*

The performance of Business English translation quality score regression is shown in Figure 3.

Figure 3 shows significant differences in the performance of different evaluation methods in the Business English translation quality score regression task. While the traditional unimodal text-based BLEURT and COMET methods can effectively capture semantic information, their errors are still relatively high. BLEURT achieves an RMSE of 0.12, a MAE of 0.10, and a MAPE of 18.37%, resulting in limited overall evaluation accuracy. COMET significantly improves upon BLEURT, with RMSE reduced to 0.09, MAE to 0.07, and MAPE to 12.43%. Multimodal models further outperform unimodal methods. VisualBERT achieves an RMSE and MAE of 0.08 and 0.06, respectively, with a MAPE of 10.18%. UNITER further reduces this to an RMSE of 0.07, a MAE of 0.05, and a MAPE of 9.62%, achieving high accuracy. LXMERT methods VisualBERT, achieving an RMSE of 0.08, a MAE of 0.06, and a MAPE of 11.03%. Overall, the multimodal BERT (BERT+ResNet50+TUTA) achieves the best performance, with RMSE and MAE of 0.06 and 0.04, respectively, and a MAPE of only 7.89%. This nearly halves the error of BLEURT and outperforms all other compared methods.

### *CORRELATION ANALYSIS WITH MANUAL SCORING*

This paper compares the correlation between various evaluation methods and manual MQM style ratings, using four metrics: Spearman, Pearson, Kendall, and ICC (Intraclass Correlation Coefficient) to comprehensively reflect the consistency and reliability of the methods with manual ratings. The results of the correlation analysis with manual ratings are shown in Figure 4.

The results in Figure 4 show that the traditional text-based BERTScore (BERT) and MoverScore (RoBERTa) have low correlations, with Spearman correlation coefficients of 0.62 and 0.65, Pearson correlation coefficients of 0.60 and 0.63, Kendall coefficients of only 0.45 and 0.48, and ICCs of only 0.58 and 0.61, respectively. This indicates limited con-

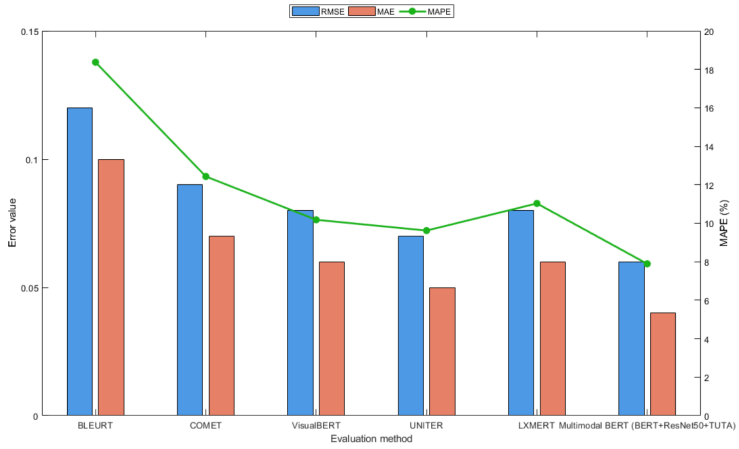


**FIGURE 3.** Business English Translation Quality Score Regression Performance

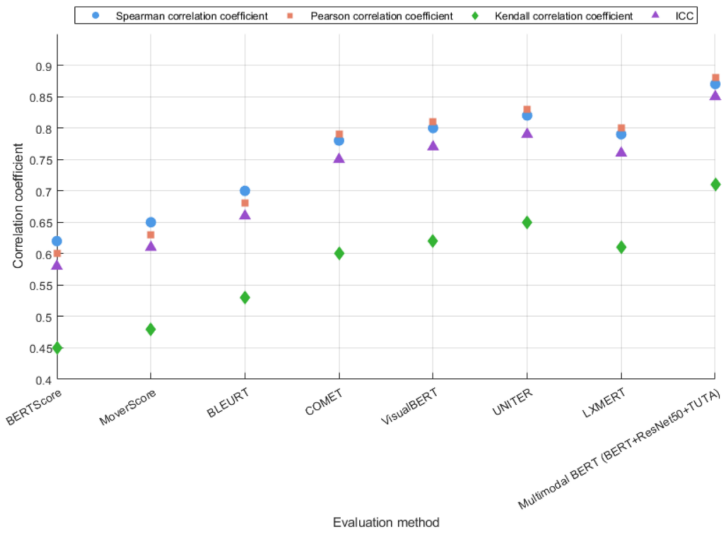


**FIGURE 4.** Correlation Analysis Results with Manual Scoring

sistency with human scorers in complex business contexts. BLEURT performs slightly better, with a Pearson correlation coefficient of 0.68 and an ICC of 0.66. COMET stands out among unimodal methods, with a Pearson correlation coefficient of 0.79, a Spearman correlation of 0.78, and an ICC of 0.75. Multimodal models incorporating visual information outperform unimodal methods overall. VisualBERT, UNITER, and LXMERT achieve Pearson correlation coefficients of 0.81, 0.83, and 0.80, respectively, and ICCs of 0.77, 0.79, and 0.76, respectively, demonstrating consistent performance. UNITER, in particular, performs particularly well across all three types of correlation coefficients. The multimodal BERT (BERT+ResNet50+TUTA) achieves the best performance, with a Spearman coefficient of 0.87, a Pearson correlation coefficient of 0.88, an improved Kendall coefficient of 0.71, and an ICC of 0.85, significantly outperforming other methods and demonstrating the highest degree of agreement with human subjective judgment in capturing multimodal semantic consistency in translation.

To further explore the significant differences in the correlation between different evaluation methods and human ratings, a paired t-test is performed. The experimental steps are as follows:

(1) For each comparison method, set the null hypothesis H0: the mean difference of the Pearson correlation coefficient between multimodal BERT and the baseline method is 0; the alternative hypothesis H1: the mean difference between the two is not 0 (two-sided test).

(2) Stratified bootstrap resampling/stability partitioning is used to obtain 30 folds for the test set. The Pearson correlation coefficient between multimodal BERT and a baseline method and human ratings is calculated on each fold, and a difference sequence is constructed.

(3) The sample mean and sample standard deviation are calculated and the test statistic is tested. A twosided p-

value is used and a 95% confidence interval is reported. The significance level is set to 0.05. The results of the significance analysis are shown in Table 2.

**TABLE 2.** Significance Analysis Results

| Comparison methods | T-value | P-value | Significance | 95% confidence interval (mean difference) |
|---|---|---|---|---|
| BERTScore(BERT) | 15.34 | < 1e-10 | Significant | (0.24, 0.32) |
| MoverScore(RoBERTa) | 15.21 | < 1e-10 | Significant | (0.22, 0.28) |
| BLEURT | 12.16 | < 1e-09 | Significant | (0.17, 0.23) |
| COMET | 7.05 | 7.8e-08 | Significant | (0.06, 0.12) |
| VisualBERT | 5.48 | 1.2e-05 | Significant | (0.04, 0.10) |
| UNITER | 4.57 | 6.5e-05 | Significant | (0.03, 0.07) |
| LXMERT | 6.27 | 3.4e-07 | Significant | (0.05, 0.11) |

The significance analysis results in Table 2 show that Multimodal BERT significantly differs from all baseline methods in Pearson correlations with human scores. Among the text unimodal baselines, BERTScore (BERT) and MoverScore (RoBERTa) show the most significant differences, with t-values of 15.34 and 15.21, respectively, and p-values less than 1e-10. The corresponding 95% confidence intervals for the mean differences are (0.24, 0.32) and (0.22, 0.28), respectively, indicating that Multimodal BERT achieves the greatest improvement. BLEURT shows the next greatest improvement, with a t-value of 12.16, p < 1e-09, and a mean difference interval of (0.17, 0.23). For the strong baseline COMET, the difference remains significant but smaller, with a t-value of 7.05, p = 7.8e-08, and a confidence interval of (0.06, 0.12), demonstrating that multimodal information remains effective in improving the upper bound. Among the multimodal baselines, the gap between VisualBERT, UNITER, and LXMERT and Multimodal BERT has narrowed, with UNITER achieving the smallest difference, with a t-value of 4.57 and a confidence interval of only (0.03, 0.07), indicating that its performance is close to that of the proposed method, but still shows a statistically significant improvement.

### *VERIFICATION OF MULTIMODAL EFFECTS*

To verify the actual contribution of each modality to Business English translation quality assessment, the experiment compares the correlation and regression performance of a single modality (text), two bimodal combinations (text + image, text + structured table), and a complete trimodal combination (text + image + structured table). Based on this, the relative contribution of each modality to overall performance is calculated. The results of the multimodal validation are shown in Figure 5.

In Figure 5(a), from the correlation coefficient, as the modalities are gradually added, the consistency between the model and the manual scoring is significantly improved. In the single-mode text mode, the Spearman correlation coefficient is 0.72, the Pearson correlation coefficient is 0.74, the Kendall correlation coefficient is 0.56, and the ICC is 0.68; after adding pictures, the Spearman correlation coefficient becomes 0.80, the Pearson correlation coefficient is 0.81, the Kendall correlation coefficient is 0.60, and the ICC is 0.77 (compared with text, they are improved by 0.08, 0.07, 0.04, and 0.09 respectively); when paired with text + table, the Spearman correlation coefficient further increases to 0.83, the Pearson correlation coefficient is 0.84, the Kendall correlation coefficient is 0.64, and the ICC is 0.80 (compared with text, they are improved by 0.11, 0.10, 0.08, and 0.12 respectively); in the complete trimodality (text + picture + structured table), the correlation coefficient reaches the highest, with Spearman correlation coefficient of 0.87, Pearson correlation coefficient of 0.88, Kendall correlation coefficient of 0.71, and ICC of 0.85, which are improved by 0.08, 0.07, 0.04, and 0.09 respectively compared with pure text. 0.15, 0.14, 0.15, and 0.17. It can be seen that both visual and tabular information can significantly improve the linear and rank correlation between the model and human ratings, and the three-modal fusion produces cumulative gains.

In Figure 5(b), regression error decreases with modality enhancement. The MAE for text is 0.07 and the RMSE is 0.10. Both the text + image and text + table bimodal combinations reduce the MAE to 0.05 and the RMSE to 0.07. The full trimodal combination further reduces the MAE to 0.04 and the RMSE to 0.06. This demonstrates that, at the numerical regression level, image and table information significantly contribute to reducing absolute prediction error, and their combined use further compresses residuals, improving scoring accuracy and stability. In Figure 5(c), breaking down the contribution by degree, text accounts for 50.34% of the total, images for 30.53%, and structured tables for 19.13%. This indicates that text remains the most critical modality, but visual information makes a significant marginal contribution to performance, and tables provide non-negligible specialized information. It can be seen that although the individual gain of text $\rightarrow$ table (Spearman +0.11) is greater than that of text $\rightarrow$ image (Spearman +0.08) in some correlation indicators, images account for a higher proportion in the overall contribution, indicating that images provide a broader and more stable gain for various types of documents. The gain of tables is more concentrated when structured numerical information appears, but the sample coverage is relatively narrow, so the overall proportion is relatively small.

### *COMPUTATIONAL EFFICIENCY AND DEPLOYMENT PERFORMANCE CONSIDERATIONS*

To evaluate the feasibility of various methods in engineering deployment, this paper compares metrics such as single-sample inference latency, GPU (Graphics Processing Unit)/CPU (Central Processing Unit) memory usage, GFLOPs (Giga Floating Point Operations Per Second), and model parameter count to reflect model real-time performance, resource consumption, and deployment cost. Table 3 shows the computational efficiency and deployment performance results.

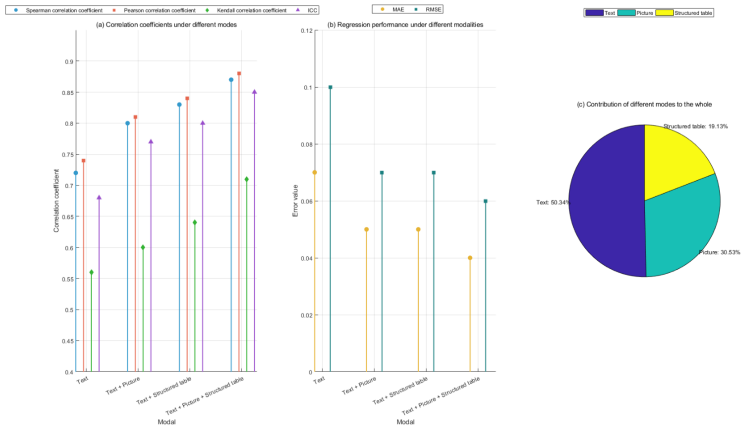


**FIGURE 5.** Multimodal effect verification results; (a) Correlation coefficients for different modes; (b) Regression performance for different modes; (c) Contribution of different modes to the overall model

**TABLE 3.** Computational Efficiency and Deployment Performance Results

| Evaluation methods | Single-sample inference time (ms) | GPU memory usage (GB) | CPU memory usage (GB) | GFLOPs | Model parameters (M) |
|---|---|---|---|---|---|
| BLEURT | 28.45 | 1.85 | 0.48 | 48.5 | 110.25 |
| COMET | 42.12 | 2.6 | 1.05 | 68.2 | 148.9 |
| VisualBERT | 89.5 | 6.78 | 2.2 | 175.4 | 179.55 |
| UNITER | 238.67 | 12.34 | 4.12 | 642.8 | 221.67 |
| LXMERT | 95.3 | 7.48 | 2.45 | 198.75 | 203.1 |
| Multimodal BERT (BERT+ResNet50+TUTA) | 86.73 | 6.12 | 2.1 | 152.3 | 132.4 |

In Table 3, in terms of computational efficiency and deployment performance, the lightweight text methods BLEURT and COMET have the shortest inference times, at 28.45 ms and 42.12 ms, respectively. They use only 1.85 GB and 2.6 GB of GPU memory, have 48.5 GFLOPs and 68.2 GFLOPs, respectively, and have only 110.25M and 148.9M parameters, demonstrating their high efficiency and low power consumption. Among the cross-modal methods, VisualBERT and LXMERT have significantly higher complexity, with inference times of 89.5 ms and 95.3 ms, respectively. The GPU occupies 6.78 GB and 7.48 GB, GFLOPs of 175.4 and 198.75, and the number of parameters also increases to 179.55M and 203.1M, significantly increasing the deployment cost. UNITER has the highest cost, with an inference time of 238.67 ms, a GPU occupancy of up to 12.34 GB, GFLOPs of 642.8, and a parameter count of 221.67M, resulting in the worst real-time performance and resource efficiency. In comparison, the multimodal BERT (BERT+ResNet50+TUTA) achieves a balance between performance and efficiency, with an inference time of 86.73 ms, a GPU footprint of 6.12 GB, a CPU footprint of 2.1 GB, 152.3 GFLOPs, and 132.4M parameters. Its overall computational overhead is lower than that of VisualBERT, LXMERT, and UNITER, while maintaining similar multimodal expression capabilities. This demonstrates its superior deployment feasibility while ensuring the accuracy of translation quality assessment.

### *ROBUSTNESS AND GENERALIZATION VERIFICATION*

Based on standard manually annotated text, it randomly injects character-level errors, such as letter substitutions, letter omissions, letter insertions, and letter position swaps. It gradually increases the error injection rate according to preset ratios (e.g., 5%, 10%, 20%, 30%, etc.) to create noise scenarios of varying intensities. The robustness and generalization verification results are shown in Figure 6. In Figure 6, COSCO stands for China Ocean Shipping Company.

As shown in Figure 6, as the intensity of character-level noise increases, the model's correlation and regression accuracy decrease. From 1% to 30% error injection, the Spearman correlation coefficient decreases from 0.86 to 0.64 (an absolute decrease of 0.22), and the Pearson correlation coefficient decreases from 0.87 to 0.69 (an absolute decrease of 0.18). Simultaneously, the regression error increases, with the MAE increasing from 0.04 to 0.12 and the RMSE from 0.06 to 0.14. A gradual degradation is observed in the intermediate stages: at 5% error injection, the Spearman correlation coefficient is 0.82, the Pearson correlation coefficient is 0.84, the MAE is 0.05, and the RMSE is 0.07; at 10% error injection, the Spearman correlation coefficient is 0.79, the Pearson correlation coefficient is 0.81, the MAE is 0.07, and the RMSE is 0.09. The overall trend shows that the impact of a small amount of character perturbation on the model is well controllable (1–5% has a small change), but the performance deteriorates to a certain extent when the noise reaches double digits.

In cross-company testing, using this company as the benchmark, the Spearman coefficient decreases to 0.75, the Pearson correlation coefficient to 0.74, the MAE to 0.08, and the RMSE to 0.11 on the China National Petroleum Corporation (CNPC) sample (an absolute decrease of 0.12 in Spearman terms and a doubling of the MAE compared to the benchmark). Performance improves significantly on Sinochem Group (Spearman coefficient of 0.83, Pearson correlation coefficient of 0.82, MAE to 0.06, and RMSE to 0.08), while maintaining near-training domain performance

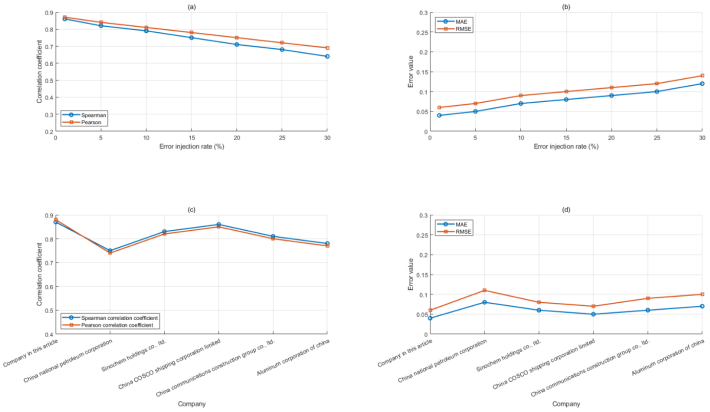


**FIGURE 6.** Robustness and generalization verification results: (a) Spearman and Pearson correlation coefficients at different error injection rates; (b) MAE and RMSE at different error injection rates; (c) Spearman and Pearson correlation coefficients for different company data; (d) MAE and RMSE for different company data

on China COSCO Shipping Corporation Limited. Spearman coefficients for China Communications Construction Group Co., Ltd. and Aluminum Corporation of China Limited are 0.81 and 0.78, respectively, with RMSEs of 0.09 and 0.10. This indicates that the model generalizes best for companies closest to the training domain or with similar document formats (such as CNOOC). For companies with dense terminology or significantly different formats (such as CNPC), both correlation and error deteriorate to some extent.

### *DIFFERENCES IN BUSINESS ENGLISH QUALITY ASSESSMENT FOR DIFFERENT DOCUMENT TYPES*

The experiment compares the consistency of various evaluation methods with human scorers on four typical business document types: financial statements, product manuals, press releases, and sustainability reports. The goal is to demonstrate the performance differences between different methods under varying document structures and visual/table dependencies. Figure 7 shows the differences in Business English quality assessment across different document types.

In Figure 7, in the financial statement task, the Spearman scores of BERTScore (BERT) and MoverScore (RoBERTa) are only 0.58 and 0.60, respectively, while BLEURT slightly improves (0.66); COMET reaches 0.76 and 0.77 in Spearman and Pearson, respectively, and UNITER 0.74 and 0.75. Multimodal BERT achieves the highest score, with Spearman of 0.84, Pearson of 0.85, Kendall of 0.68, and ICC of 0.82, an improvement of 0.08 in Pearson compared to COMET.

### *BUSINESS VALUE ASSESSMENT*

To evaluate the potential value returns of each method in actual business scenarios, four business indicators, namely, labor hour savings rate, avoidable economic risk reduction ratio, process automation rate improvement, and customer complaint rate, are edited to quantify the model's contribution to operational efficiency and risk control. The results are shown in Figure 8.

In Figure 8, looking at the labor-hour savings and process automation rates, BERTScore (BERT) achieves a labor-hour savings rate of only 12.34% and an automation rate increase of 10.22%. BLEURT shows some improvement, achieving 21.45% and 17.89%, respectively. With enhanced semantic modeling capabilities, COMET increases its labor-hour savings to 28.30% and its automation rate to 25.10%. Multimodal methods perform even better, with UNITER achieving a labor-hour savings rate of 36.44% and an automation rate of 37.12%. The best approach is multimodal BERT (BERT+ResNet50+TUTA), achieving a labor-hour savings rate of 49.83% and an automation rate increase of 54.11%. These improvements, respectively, represent 37.49% and 43.89% increases over BERTScore (BERT), demonstrating its advantages in significantly reducing labor input and significantly improving process automation.

In terms of risk control and customer experience, BERTScore (BERT) only reduces economic risk by 6.78%, while the customer complaint rate remains high at 11.45%. MoverScore (RoBERTa) and BLEURT improve to 8.90% and 12.34% risk reduction, respectively, with complaint rates of 10.12% and 8.76%. COMET showed significant improvement, reducing economic risk by 18.45% and its complaint rate to 6.95%. VisualBERT and LXMERT further improved, reaching 22.78% and 23.50% risk reduction, respectively, with complaint rates of 5.60% and 5.20%. UNITER even achieved a higher risk reduction rate of 25.63% and a complaint rate of only 4.98%. The best-performing multimodal BERT not only increases the economic risk reduction ratio to 38.64% but also reduces the customer complaint rate to 3.21%, a decrease of 5.55% compared to BLEURT's 8.76% complaint rate, highlighting its business value in reducing potential losses and improving customer satisfaction.

## V. CONCLUSION

This paper develops a Multimodal BERT-based quality estimation framework specifically designed for complex business scenarios. BERT encodes subword-level context

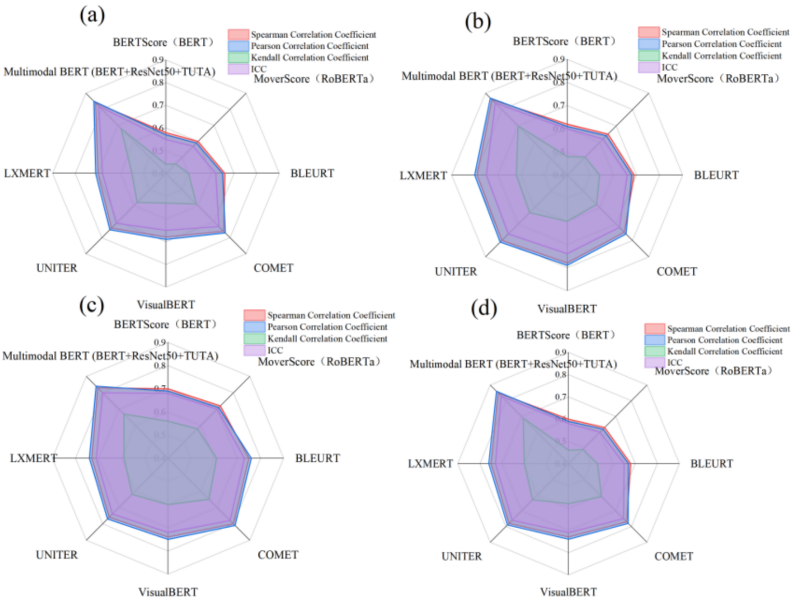


**FIGURE 7.** Differences in business English quality assessments by document type: (a) Financial statement; (b) Product brochure; (c) Press release; (d) Sustainability report

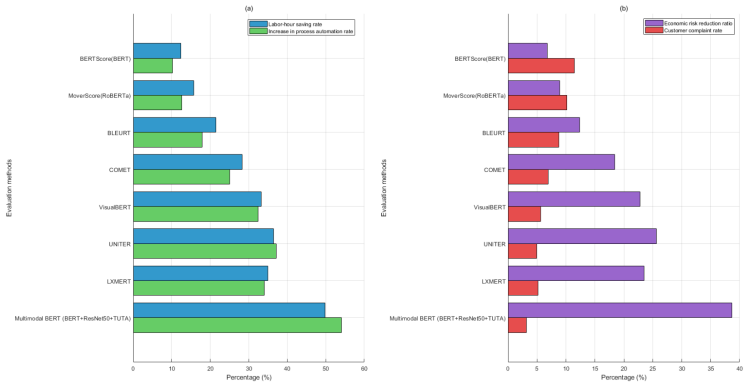


**FIGURE 8.** Business value assessment: (a) Man-hour savings and process automation rate increase; (b) Economic risk reduction and customer complaint rate

for source and translation text, a ResNet50 model extracts multi-scale image features, and TUTA encodes the semantics of rows, columns, and cells in tables. The three modalities are then projected into a shared space via linear projection. Multi-head cross-modal attention is used for interactive alignment of text-image and text-table. Gated weighted pooling is then used to generate a fused vector. Finally, two fully connected regression heads concurrently optimize the WMSE, modality alignment, and ranking losses to output a normalized quality score. Experimental results show that this method achieves an RMSE of 0.06 on sentence-level regression tasks, with Spearman and Pearson correlations of 0.87 and 0.88 with manual MQM, respectively. The Pearson coefficient improves by 0.20 compared to BLEURT. This method also delivers significant business benefits (38.64% of financial risk avoided and a 3.21% customer complaint rate). The shortcomings are that it is still sensitive to severe OCR noise and cross-industry terminology distribution shifts, and the model calculation and deployment costs need to be further optimized. Future work can focus on lightweighting, multilingual expansion, and enhancing robustness to OCR noise, all of which are critical for real-world application.

## AUTHOR CONTRIBUTIONS

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

## CONFLICTS OF INTEREST

The author declares no conflict of interest.

## FUNDING

This research received no external funding.

## ACKNOWLEDGEMENTS

Not applicable.

## REFERENCES

[1] D. Tang, "Study on the English-Chinese Translation of Business Contracts from the Perspective of Skopos Theory," Journal of Theory and Practice in Linguistics, vol. 1, no. 3, pp. 1-9, 2024.

[2] S. Liu, Y. Chen, K. Xu, and J. Lin, "Emotional analysis of evaluation discourse in business English translation based on language big data mining of public health environment," Frontiers in Public Health, vol. 10, Art. no. 981182, 2022, doi: 10.3389/FPUBH.2022.981182.

[3] V. Agustina, R. Thamrin N, and E. Oktoma, "The role of English language proficiency in the global economy and business communication," International Journal Administration, Business & Organization, vol. 5, no. 4, pp. 82-90, 2024, doi: 10.61242/ijabo.24.423.

[4] G. Tang, O. Yousuf, and Z. Jin, "Improving BERTScore for machine translation evaluation through contrastive learning," IEEE Access, vol. 12, no. 1, pp. 77739-77749, 2024, doi: 10.1109/ACCESS.2024.3406993.

[5] D. Beauchemin, H. Saggion, and R. Khoury, "MeaningBERT: assessing meaning preservation between sentences," Frontiers in Artificial Intelligence, vol. 6, no. 6, Art. no. 1223924, 2023, doi: 10.3389/frai.2023.1223924.

[6] C. Han, "Quality assessment in multilingual, multimodal, and multiagent translation and interpreting (QAM3 T&I): Proposing a unifying framework for research," Interpreting and Society, vol. 5, no. 1, pp. 27-55, 2025, doi: 10.1177/27523810251322645.

[7] T. Tayir and L. Li, "Unsupervised multimodal machine translation for low-resource distant language pairs," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 4, pp. 1-22, 2024, doi: 10.18653/v1/2024.findings-emnlp.320.

[8] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," Artificial Intelligence Review, pp. 58(6):, 2025, doi: 1-49.10.1007/s10462-025-11162-5.

[9] Y. Cui and M. Liang, "Automated scoring of translations with BERT models: Chinese and English language case study," Applied Sciences, vol. 14, no. 5, pp. 1925-1941, 2024, doi: 10.3390/APP14051925.

[10] Z. Xiao, X. Ning, and M. J. M. Duritan, "BERT-SVM: A hybrid BERT and SVM method for semantic similarity matching evaluation of paired short texts in English teaching," Alexandria Engineering Journal, vol. 126, no. 1, pp. 231-246, 2025, doi: 10.1016/J.AEJ.2025.04.061.

[11] D. Guo, "Deep learning-driven context-aware English translation for ambiguous sentences," International Journal of Information and Communication Technology, vol. 26, no. 15, pp. 41-56, 2025, doi: 10.1504/IJICT.2025.146373.

[12] M. Alamri and S. Lajmi, "Design a smart platform translating Arabic sign language to English language," Int. J. Electr. Comput. Eng. IJECE, vol. 14, no. 4, pp. 4759-4774, 2024, doi: 10.11591/ijece.v14i4.pp4759-4774.

[13] V. D. Avina, M. Amiruzzaman, S. Amiruzzaman, L. B. Ngo, and M. A. A. Dewan, "An AI-Based Framework for Translating American Sign Language to English and Vice Versa," Information, vol. 14, no. 10, pp. 569-579, 2023, doi: 10.3390/info14100569.

[14] M. Roweida, A. Inad, M. Al-Ayyoub, and A. Fadel, "Multimodal Multi-source Neural Machine Translation: Building Resources for Image Caption Translation from European Languages into Arabic," Computation, vol. 13, no. 8, pp. 194-204, 2025, doi: 10.3390/computation13080194.

[15] E. Tian, Z. Zhu, F. Liu, Z. Li, R. Gu, and S. Zhao, "Multimodal Machine Translation Based on Enhanced Knowledge Distillation and Feature Fusion," Electronics, vol. 13, no. 15, pp. 3084-3099, 2024, doi: 10.32604/cmc.2025.061145.

[16] H. Yu, R. Ma, M. Su, P. An, and K. Li, "A novel deep translated attention hashing for cross-modal retrieval," Multimedia Tools and Applications, vol. 81, no. 18, pp. 26443-26461, 2022, doi: 10.1007/s11042-022-12860-w.

[17] R. Cai, J. Dong, T. Liang, Y. Liang, Y. Wang, X. Yang, et al., "Cross-lingual cross-modal retrieval with noise-robust fine-tuning," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 11, pp. 5860-5873, 2024, doi: 10.1109/TKDE.2024.3400060.

[18] A. Shin, M. Ishii, and T. Narihira, "Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision," International journal of computer vision, vol. 130, no. 2, pp. 435-454, 2022, doi: 10.1007/s11263-021-01547-8.

[19] J. Guo, J. Ye, Y. Xiang, and Z. Yu, "Layer-level progressive transformer with modality difference awareness for multi-modal neural machine translation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, no. 1, pp. 3015-3026, 2023, doi: 10.1109/TASLP.2023.3301210.

[20] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Computing and Applications, vol. 35, no. 23, pp. 17095-17112, 2023, doi: 10.1007/S00521-023-08568-Z.

[21] I. V. Modoranu, M. Safaryan, G. Malinovsky, E. Kurtic, T. Robert, P. Richtarik, et al., "Microadam: Accurate adaptive optimization with low space overhead and provable convergence," Advances in Neural Information Processing Systems, vol. 37, no. 1, pp. 1-43, 2024.

[22] C. Zhang, Y. Shao, H. Sun, L. Xing, Q. Zhao, and L. Zhang, "The WuC-Adam algorithm based on joint improvement of Warmup and cosine annealing algorithms," Mathematical Biosciences & Engineering, vol. 21, no. 1, pp. 1270-1285, 2024, doi: 10.3934/mbe.2024054.

[23] O. V. Johnson, C. Xinying, K. W. Khaw, and M. H. Lee, "ps-CALR: periodic-shift cosine annealing learning rate for deep neural networks," IEEE Access, vol. 11, no. 1, pp. 139171-139186, 2023.