ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Multi-modal English Translation Production Model Combining Cross-modal Alignment and Attention Mechanism

Shufang Wang[1,*]

[1] School of Foreign Languages, Zhengzhou Shengda University of Economics and Management, Zhengzhou 451191, Henan, China

SUMMARY: *Aiming at the problems of traditional English translation models, such as insufficient scene constraints, limited ambiguity resolution ability, and unstable image-text semantic coordination, this paper constructed a multimodal English translation production model combining cross-modal alignment and attention mechanism. Based on the collaborative input of text and image, the model forms an integrated technical link of "alignment-attention-generation" through multimodal input representation, shared semantic space mapping, bidirectional cross-modal alignment and attention-driven decoding generation. Experimental results show that the BLEU, METEOR and ROUGE-L of the proposed model on the test set reach 37.4, 32.5 and 41.3 respectively, which are 5.6, 4.4 and 5.9 percentage points higher than those of the basic Transformer model. The accuracy of image-text consistency, ambiguity resolution and entity alignment reaches 85.9%, 84.2% and 85.1%, respectively. The results show that cross-modal alignment can effectively reduce the representation deviation between text semantics and visual semantics, and the attention mechanism can enhance the dynamic screening ability of key contexts in the translation generation stage, thereby improving the accuracy, stability and application adaptability of multimodal English translation production.*

KEYWORDS: *Multimodal English translation; Cross-modal alignment; Attention mechanism; Translation production model*

1 Introduction

1.1 Research Background on Multimodal English translation production

With the transformation of artificial intelligence systems from single text processing to collaborative understanding of multi-source information such as text, image, and voice, the research paradigm of translation tasks is also changing significantly. Yuan et al. (2025) systematically reviewed the methods, applications and future trends of multimodal learning, and pointed out that cross-modal information modeling, heterogeneous feature fusion and unified representation learning have become the key directions for the development of multimodal intelligence [1]. This change means that English translation production is no longer just a linear mapping from source text to target text, but gradually evolves into a composite process of context completion, ambiguity resolution and target generation relying on multi-source semantic clues. Jin et al. (2025) further pointed out that multi-modal large language models have shown strong capabilities in visual question answering, visual

*101048@shegnda.edu.cn
https://doi.org/10.65102/is20261051 Published: 10 June, 2026

understanding and reasoning tasks, but the high training cost and high inference overhead still limit their widespread deployment [2]. Therefore, exploring multimodal models with both performance improvement and engineering realizability in translation production scenarios has become an important issue in the cross research of computer technology and language intelligence.

Focusing on the task of English translation production, the existing research has gradually expanded from traditional neural machine translation to vision-assisted translation, multi-source translation and retrieval enhanced translation. Mohammed et al. (2025) proposed a multi-modal multi-source neural machine translation resource construction scheme, emphasizing that the combination of text and visual input is helpful to improve the translation accuracy, and multi-source input is especially supportive for low-resource target language scenes [3]. Tian et al. (2024) proposed a multimodal machine translation model based on knowledge distillation and feature refinement. Aiming at the problem that image information is not directly available in the inference stage, distillation and feature compensation are used to enhance the practical usability of the model [4]. Guo et al. (2024) proposed a multi-granularity visual pivot-guided multimodal neural machine translation method, and introduced a text-aware cross-modal contrast decoupling strategy to alleviate the difference between language representation and visual representation [5]. Li (2025) proposed a multi-modal retrieval enhanced generative translation framework, which couples text information, visual information and external retrieval information into the same generative link, expanding the knowledge source of the translation model [6]. Shi et al. (2025), from the perspective of the consistency of source language, target language and visual information, introduce visual features at both the encoder and decoder to enhance the context coordination ability in the translation process [7]. The above studies show that multimodal English translation production has developed from the primary form of "image-aided translation" to a complex computational task that integrates cross-modal semantic mapping, dynamic information selection and multi-source generative control.

1.2 Research Status of Cross-modal Alignment and Attention Mechanism

Cross-modal alignment is the core premise of multi-modal English translation production. Its goal is to map textual information and visual information into a comparable and interactive unified semantic space, so as to improve the model's ability to grasp the scene meaning, anaphora relationship and context constraints. Liu et al. (2024) proposed a joint modeling method of bilingual-visual consistency and target language-visual consistency to explicitly strengthen the correspondence between source language, target language and visual annotations in multimodal neural machine translation, and extract information related to future target contexts from visual annotations with the help of the attention layer [8]. Guo et al. (2024) proposed the progressive modal complementary aggregation MultiTransformer, which gradually Narrows the gap between modalities and captures domain-related information to improve the feature collaboration ability in domain multimodal translation [9]. Zhu et al. (2024) proposed VisTFC, a Vision-guided target-side future context learning framework, which further uses the attention mechanism to generate constraints on the target side, so that visual information not only serves the encoding stage, but also participates in the context prediction of the translation generation stage [10]. These studies show that cross-modal alignment has shifted from the early coarse-grained stitching to a research path that emphasizes more fine-grained consistency, target-side constraints, and temporal generation synergy.

The attention mechanism provides dynamic control capabilities for information screening,

weight allocation and key semantic focus in multimodal translation. Fang et al. (2025) proposed a scalable multimodal representation learning network, which maps multimodal features into a shared representation space by learning a mode-specific projection matrix to improve high-order information preservation and out-of-sample generalization [11]. Wei et al. (2025) proposed a knowledge-enhanced vision-language contrastive representation learning framework to improve the alignment quality between visual features and language features with the help of domain knowledge [12]. Yao et al. (2024) proposed a multi-scale visual attention network driven by language conditions, which uses both visual information and language information to carry out multimodal reasoning in the visual positioning task, reflecting the fine advantage of attention mechanism in cross-modal correlation modeling [13]. Zhao et al. (2025) reformulated multi-modal entity alignment as an entailment judgment problem, and proposed a multi-modal entity entailment framework to achieve semantic matching between heterogeneous information in a more discriminative way [14]. Jimenez-Guarneros and Fuentes-Pineda (2025) proposed a multi-level alignment and consistent decision boundary method to achieve more robust knowledge transfer in cross-agent transfer scenarios, indicating that cross-modal alignment research is gradually moving from static feature mapping to a unified design that considers both hierarchical structure and robustness [15]. In general, the existing research has fully explained the important role of cross-modal alignment and attention mechanism for multi-modal modeling. However, there is still room for further deepening how to truly integrate the two mechanisms into the complete link of "input representation - alignment modeling - translation generation - result optimization" in English translation production scenarios.

1.3 Research objectives and innovations

Based on the above research background and current situation, this paper intends to construct a multimodal English translation production model combining cross-modal alignment and attention mechanism. This study aims to design a unified cross-modal semantic alignment module and a dynamic attention allocation mechanism for text-visual collaborative input scenarios, which can reduce the modal semantic deviation and enhance the model's ability to capture key translation cues, and achieve higher quality, stronger consistency and better robustness of English translation generation. Compared with the existing research, the innovative considerations of this paper mainly focus on three levels. First, the cross-modal alignment and attention mechanism are put into the same translation production framework for collaborative modeling to avoid the separation of them. Second, the linkage relationship between source language, visual cues and target language generation is strengthened, so that the attention mechanism is not only involved in local semantic focusing, but also serves the global cross-modal semantic routing. Thirdly, from the perspective of computer technology implementation, feature representation, semantic mapping, dynamic weighting and sequence generation are organized into a trainable, verifiable and extensible engineering process. In the past two years, related research has promoted this field from the directions of coarse-grained alignment, multi-scale task learning, cross-modal semantic fusion and multimodal large-model cognitive evaluation. Some representative literature pairs are shown in Table 1.

Table 1: Comparison of some representative literatures

Reference	Core Method	Research Focus	Implications for This Study
Guo et al. (2024) [16]	Coarse-to-fine-grained multimodal contrastive alignment network	Strengthens multi-level cross-modal alignment capability	Provides a reference for the design of the fine-grained alignment module in this study
Lin et al. (2024) [17]	Multi-scale feature extraction and multi-task learning	Improves heterogeneous modal feature fusion and task collaboration capability	Can be used to improve the hierarchical representation mechanism in translation generation
Zhang et al. (2025) [18]	Cross-modal alignment-driven semantic fusion network	Explores image-text semantic relationships in depth and corrects feature misalignment	Supports the cross-modal fusion and semantic calibration in this study
Liu et al. (2025) [19]	AU-guided visual-language alignment	Enhances the modeling capability of fine-grained visual-language relationships	Can be used to strengthen the constraint effect of visual cues on translation generation
Schulze Buschoff et al. (2025) [20]	Visual cognition evaluation of multimodal large models	Examines the boundaries of vision-language models in complex cognitive tasks	Suggests that this study should pay attention not only to performance but also to interpretability and reliability

The studies listed in Table 1 show that although the current related work has made progress in cross-modal alignment, semantic fusion and cognitive evaluation, it is still rare to directly build an integrated "alignment-attention-generation" model for the whole process of English translation production, which also forms the realistic basis for further research in this paper.

2 Construction of Multimodal English Translation Production Model Combining Cross-modal Alignment and Attention Mechanism

2.1 The overall framework design of the model

Aiming at the task of multi-modal English translation production, this paper constructs a unified translation framework that integrates cross-modal alignment mechanism and attention enhanced generation mechanism. The model takes the source language text and visual information as the main input, and completes the end-to-end mapping from heterogeneous modal input to target translation generation through five core links: text encoding, visual encoding, cross-modal semantic alignment, dynamic attention fusion and target language decoding. Compared with the traditional neural machine translation models that only rely on text sequence modeling, the proposed model introduces visual context information in the encoding stage, establishes a cross-modal unified semantic space in the intermediate

representation stage, and dynamically allocates the contribution weights of different modalities through the hierarchical attention mechanism in the decoding stage, thereby improving the recognition ability of the model for scene constraints, anreference relations and semantic ambiguity. The overall framework is shown in Figure 1.

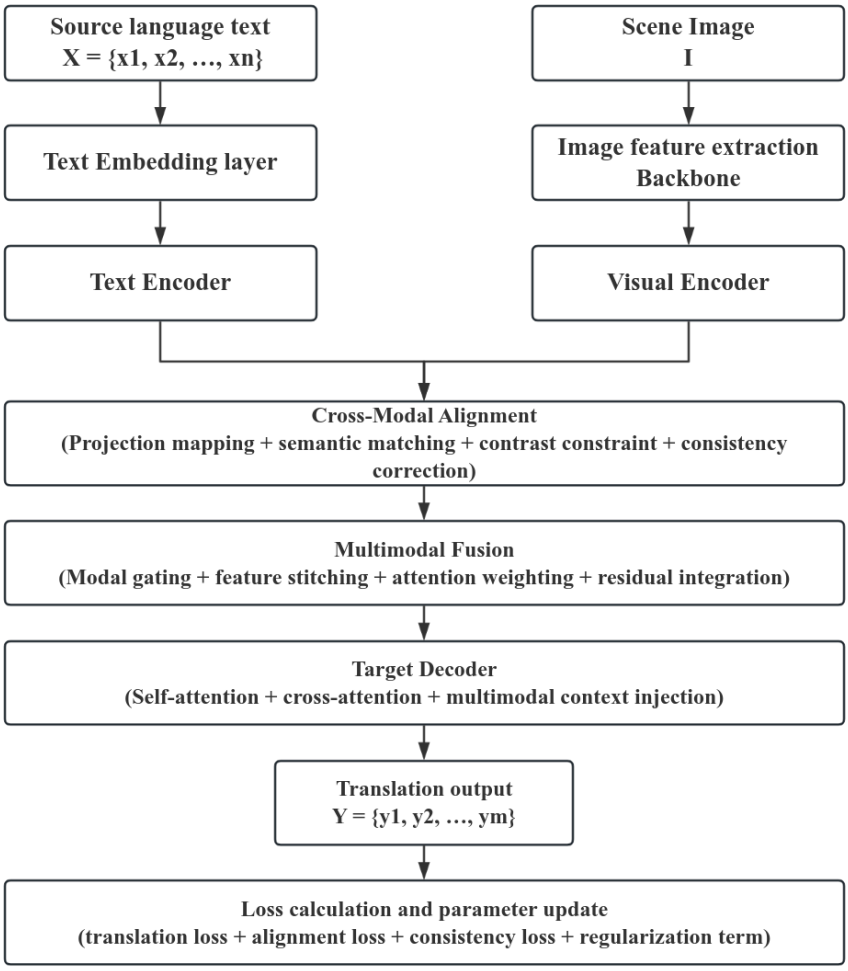

Figure 1: Design of the overall framework of the model

From the architectural point of view, the model can be regarded as a multi-stage translation production architecture of "dual encoder-alignment fusion-enhanced decoding". The input terminal receives the source text sequence and image modality information, where the source text is used to provide explicit language semantics, and the image information is used to supplement the scene content, entity relations and context constraints. Firstly, the text modality is processed by word vector mapping and position encoding, and then sent to the text encoder to extract the context semantic representation. The visual modality extracts region-level visual features through a convolutional neural network or vision Transformer, and retains the spatial semantic structure to avoid excessive compression of image information in the global pooling process. The cross-modal alignment module then maps the textual representation and visual representation into a shared semantic space, and uses the semantic correlation matrix to complete fine-grained modal matching. The aligned features enter the multi-modal fusion layer, and the feature reorganization is completed by the gated unit and the attention mechanism, and finally sent to the target language decoder for

translation generation. This structure not only retains the sequence generation advantage of neural machine translation model, but also strengthens the collaborative expression ability between different modalities, as shown in Figure 1.

Let the source text sequence be denoted as follows:

$$X = \{x_1, x_2, \ldots, x_n\} \tag{1}$$

Here, x_i denotes the ith source word. After embedding mapping and text encoder, the text context representation can be obtained as follows:

$$H^t = \text{Encoder}_t(\text{Embed}(X) + P) \tag{2}$$

Here, Embed(·) represents the word embedding function, P represents the position encoding matrix, and Ht∈Rn×d is the text feature representation.

For the visual modality, the input image I is extracted by the visual backbone network to form a region-level visual feature set:

$$H^v = \text{Encoder}_v(I) = \{v_1, v_2, \ldots, v_k\}, \quad H^v \in \mathbb{R}^{k \times d} \tag{3}$$

Here, *vj* represents the feature vector of the JTH visual region and *k* is the number of visual regions. In order to ensure the effectiveness of subsequent cross-modal interaction, this paper projects text features and visual features into the same dimensional space, and constructs a cross-modal similarity matrix:

$$S = \text{softmax}\left(\frac{(H^t W_t)(H^v W_v)^\top}{\sqrt{d}}\right) \tag{4}$$

Here, W_t and W_v are text projection matrices and visual projection matrices, respectively, and S is used to characterize the matching strength between word-level semantics and region-level visual information. This formula essentially realizes the core calculation process of cross-modal semantic alignment, and the result will be used as an important basis for the subsequent multi-modal attention allocation in the fusion layer and decoder.

In the fusion stage, this paper does not directly concatenate the two modalities, but introduces a gated control function to dynamically adjust the contribution ratio of text information and visual information. The fused representation is defined as follows.

$$H^m = G \odot \tilde{H}^t + (1 - G) \odot \tilde{H}^v \tag{5}$$

Here, $\tilde{H}^t$ and $\tilde{H}^v$ represent the aligned textual and visual features, G is the gated vector, and ⊙, represents element-wise multiplication. This mechanism can automatically determine whether the current translation process depends on the linguistic context or the visual context according to the input content, thereby reducing the interference of irrelevant modal noise on the generation of translation.

Autoregressive decoding is used in the target language generation stage. At the TTH time step, the decoder generates the current hidden state based on the historical translation sequence y<t and the multi-modal fusion context Hm:

$$s_t = \text{Decoder}(y_{<t}, H^m) \tag{6}$$

Furthermore, the generation probability of the current term can be expressed as follows.

$$P(y_t|y_{<t}, X, I) = \text{softmax}(W_o s_t + b_o) \quad (7)$$

where W_o and b_o are the output layer weights and bias terms, respectively. Equation (7) shows that the prediction of the target word is not only dependent on the previous translation sequence, but is completed on the unified semantic representation after cross-modal alignment and attention fusion, which makes the model have stronger generative ability when dealing with translation samples containing visual ambiguity, scene dependence and entity disambiguation requirements.

In order to realize multi-objective constraints in the model training process, this paper considers both translation loss and cross-modal alignment loss in the overall loss function, which is defined as follows:

$$\mathcal{L} = \mathcal{L}_{trans} + \lambda \mathcal{L}_{align} \quad (8)$$

Here, L_{trans} is the cross-entropy loss of the target word lemma prediction, L_{align} is the cross-modal semantic alignment loss, and λ is the balance coefficient. Through this joint optimization method, the model can not only improve the quality of the translation language generation, but also improve the degree of semantic coupling between different modalities.

From the perspective of engineering implementation path, the overall framework of the model in this paper has clear module boundary and strong scalability. The text Encoder can adopt BiLSTM, Transformer Encoder, or pre-trained language model structure. Vision encoders can adopt ResNet, Swin Transformer, or ViT architectures; The cross-modal alignment module can be further extended to contrastive learning, co-attention or bidirectional interactive encoding structures. The Decoder part can be combined with the standard Transformer Decoder to complete the multi-level target generation control. Therefore, the framework is not only suitable for English translation production tasks, but also provides a unified technical basis for subsequent multimodal machine translation optimization, cross-language content generation and intelligent translation assistance system design.

2.2 Multimodal input representation with cross-modal alignment methods

The key difficulty of multi-modal English translation production is not to introduce two types of information, text and image, but to map heterogeneous modalities with different sources, structures and semantic densities into a comparable and interactive unified representation space. The source text takes discrete word unit sequence as the basic carrier, emphasizing word order dependence, syntactic structure and contextual semantics. The visual modality mainly focuses on image regions, target entities and spatial relations, and pays more attention to scene constraints, entity distribution and local semantic association. If the two types of features are directly spliced, it is not only difficult to establish a stable semantic correspondence, but also may magnify the inter-modal noise and distribution deviation, which will affect the quality of translation generation. Based on this, a multi-modal processing link of "input representation - shared mapping - bidirectional alignment - gated fusion" is constructed to establish a fine-grained alignment relationship between text semantics and visual semantics. The core alignment module structure is shown in Figure 2.

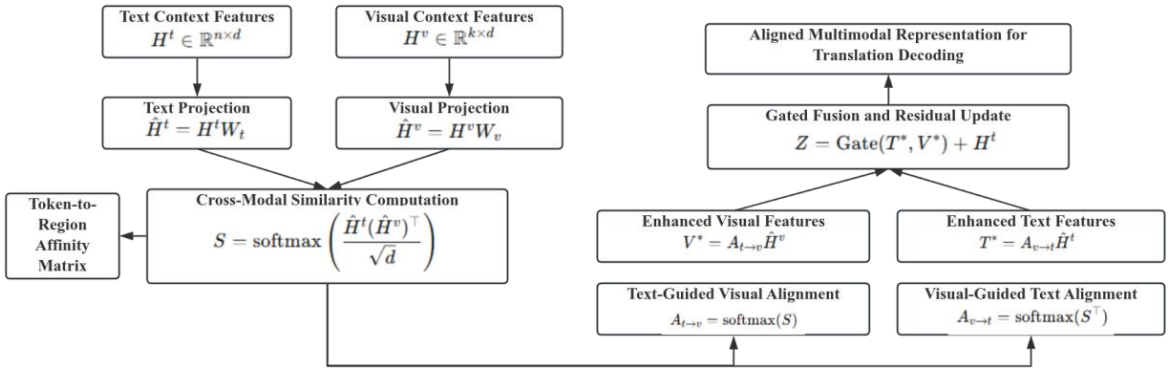

Figure 2: Detailed Structure of the Bidirectional Cross-Modal Alignment Module

As shown in Figure 2, the bidirectional cross-modal alignment module takes as input the textual context features Ht and visual context features Hv output from the previous layer. The two types of features are first mapped to a unified semantic space through an independent projection layer, and then the cross-modal similarity matrix is used to explicitly describe the matching strength between the semantics of words and visual regions. On this basis, the model performs text-guided visual alignment and visual-guided text alignment, respectively, and outputs the aligned multimodal representation through gated fusion and residual update. The function of the module is not simply superimposed text information and visual information, but to complete semantic screening, relationship correction and modal reorganization in the shared semantic space, so as to provide more stable context input for the subsequent decoding stage.

In the input representation layer, suppose that the source sentence is composed of n lemons, and its discrete input sequence is expressed as follows:

$$X = \{x_1, x_2, \dots, x_n\} \tag{9}$$

Here, xi denotes the ith source word. After embedding matrix mapping and position encoding of the text modality, the initial text representation is obtained as follows:

$$E_t = \mathrm{Embed}(X) + P_t \tag{10}$$

Here, Embed(·) represents the word vector embedding operation, and Pt is the position encoding matrix. On the one hand, the distributed semantics of the word itself is preserved, on the other hand, the intra-sentence order information is preserved, so that the model can learn the context dependencies in the subsequent stages.

The input of visual modality is not a single global image vector, but the image is divided into several regions or visual blocks to preserve local objects, entity relations and spatial distribution characteristics. Let the set of image regions be:

$$R = \{r_1, r_2, \dots, r_k\} \tag{11}$$

Here, rj denotes the JTH visual region. After the image is extracted by the visual backbone network, the initial visual representation is obtained by linear projection and spatial coding:

$$E_v = W_p \cdot \mathrm{Backbone}(R) + P_v \tag{12}$$

Here, Wp is the visual feature projection matrix and Pv is the spatial position encoding.

Different from textual position coding, which mainly describes the sequence order, visual position coding pays more attention to the relative position relationship and spatial layout structure between image regions. The multimodal input representation process is shown in Figure 3.

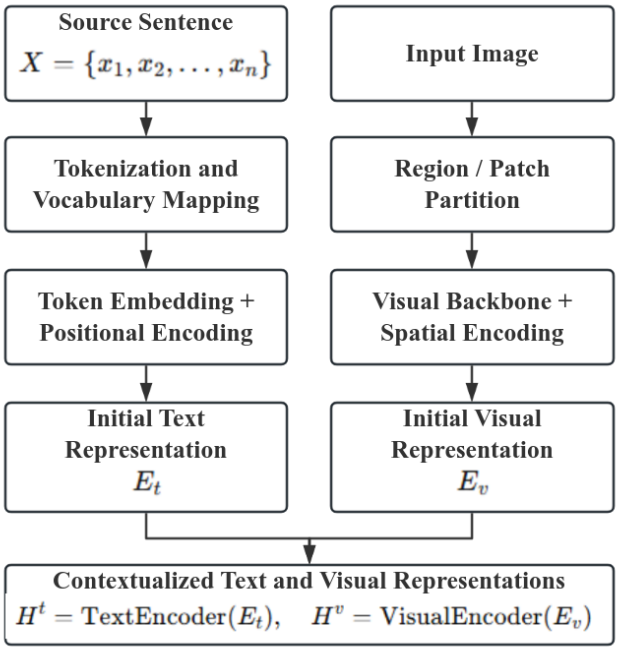

Figure 3: Multimodal Input Representation Process

After obtaining the initial representation, the intra-modal modeling is carried out by the text context encoder and the visual context encoder respectively. The text encoder is used to extract syntactic dependencies, long-distance semantic relations and context conditions, and the visual encoder is used to extract inter-region co-occurrence relations, object saliency and scene structure information. The corresponding context representation can be written as follows:

$$H^t = \text{TextEncoder}(E_t) \tag{13}$$

$$H^v = \text{VisualEncoder}(E_v) \tag{14}$$

where $H^t \in \mathbb{R}^{n \times d}, H^v \in \mathbb{R}^{k \times d}$, d denotes the hidden layer dimension. After this stage, both text and visual features have strong intra-modal expression ability, but they are still located in different semantic distribution Spaces, so they cannot directly complete effective interaction.

In order to reduce the scale deviation and representation misalignment between text semantics and visual semantics, we introduce a shared semantic projection mechanism before cross-modal interaction, and map the two types of context features into a unified latent space:

$$\overset{t}{\widetilde{H}} = H^t W_t, \quad \overset{v}{\widetilde{H}} = H^v W_v \tag{15}$$

Here, Wt and Wv represent the projection matrix of text modality and visual modality, respectively. After this process, the two types of features remain consistent in dimensional structure and semantic scale, thus providing a stable basis for subsequent similarity

calculation and attention interaction. The shared mapping is not only a formal dimension unification, but also its essence is to compress irrelevant noise and strengthen cross-modal common semantics through learnable parameters, which makes the matching between different modalities more reliable.

In the shared semantic space, the model further constructs a cross-modal similarity matrix to describe the fine-grained correspondence between the word representation and the visual region representation. The similarity matrix is defined as follows:

$$S = \text{softmax}\left(\frac{\overset{t}{\widetilde{H}}\,(\overset{v}{\widetilde{H}})^{\top}}{\sqrt{d}}\right) \tag{16}$$

Here, S∈Rn×k, and the matrix element sij represents the semantic matching strength between the ith lemma and the JTH visual region. Equation (16) uses the scaled dot product form to depict the degree of association between text semantics and visual semantics, so that the model can identify key matching relations at the word level and the region level.

Based on the similarity matrix, this paper constructs a bidirectional cross-modal alignment mechanism. Text-guided visual alignment takes text semantics as the query condition, and filters the most relevant local information from the visual area to the current word sense or phrase meaning, which is calculated as follows:

$$V^{*} = S\,\overset{v}{\widetilde{H}} \tag{17}$$

Here, V∗ represents the visually enhanced representation after text constraint. This process can effectively suppress the interference of background regions and irrelevant regions, so that visual information is no longer involved in the translation in the form of the overall scene, but in the selective context oriented to specific language semantics.

In contrast, Vision-guided text alignment uses scene information to reverse correct and complement the text representation, which is calculated as follows:

$$T^{*} = S^{\top}\,\overset{t}{\widetilde{H}} \tag{18}$$

Here, T∗ represents the text enhanced representation after visual constraints. When there are semantic ambiguities, ambiguous entity references or unclear action relations in the source language, visual information can provide additional scene evidence to enhance the discriminability of text representation. Compared with the methods that only use one-way cross-modal attention, the bidirectional alignment mechanism retains the constraint ability of both "text query vision" and "visual correction of text meaning", so it is more suitable for semantic disambiguation and context completion in multimodal translation tasks.

Considering that the text enhanced representation T∗ and the visual enhanced representation V∗ after bi-directional alignment still have differences in information quality and noise distribution, we further introduce a gated fusion mechanism to adaptively integrate the two types of alignment results. The gating vector is defined as follows:

$$g = \sigma(T^{*}W_{g} + V^{*}U_{g} + b_{g}) \tag{19}$$

where σ(·) is the Sigmoid activation function and Wg, Ug and bg are learnable parameters. Based on the gated values, the aligned multimodal representation of the final output is written

as follows:

$$Z = g \odot T^* + (1 - g) \odot V^* + H^t \tag{20}$$

Here, ⊙ denotes element-wise multiplication and Z is the unified representation fed into the subsequent translation decoder. In this equation, the gating term is used to dynamically balance the contribution ratio of text information and visual information, and the residual term Ht is used to retain the semantic backbone of the original text to avoid excessive correction or semantic drift in the process of cross-modal interaction. For the input with clear semantics and little dependence on scene information, the model will automatically increase the proportion of text features. For inputs that are highly dependent on image content for disambiguation, the model will increase the weight of visual features.

From the perspective of computer technology implementation, the multi-modal input representation and cross-modal alignment method in this paper have a clear structure layering and strong engineering scalability. In the input representation stage, a modal independent coding strategy is adopted, which can give full play to the representation advantages of text encoders and visual encoders in their respective fields. The shared semantic projection stage explicitly alleviates the scale misalignment problem of heterogeneous features. In the bidirectional alignment stage, fine-grained semantic relationship is constructed by similarity matrix and bidirectional attention. In the output stage, gated fusion and residual update are used to control the information flow, so as to improve the stability and robustness of the model in complex scenes. Overall, this method provides a higher quality multimodal semantic input for the subsequent attention-based translation generation process, and also lays a method foundation for scene constraint, entity disambiguation and context consistency control in English translation production.

2.3 Attention-driven English translation production process

After the multi-modal input representation and cross-modal alignment are completed, the model also needs to effectively inject the aligned semantic information into the target language generation process to achieve dynamic translation construction for English translation production tasks. Different from traditional neural machine translation which only relies on the source language context for decoding, this paper introduces an attention mechanism in the translation generation stage, so that the decoder can adaptively select the most relevant text semantics, visual semantics and their alignment results according to the current generation state at each time step, thereby improving the accuracy, context consistency and semantic interpretability of the translation. The core of this process is not to simply read multimodal features, but to complete the dynamic association between "decoding state, cross-modal representation and word prediction" through attention allocation.

From the perspective of the generation process, attention-driven English translation production mainly includes three key steps. The first step is to construct the decoding state of the current time step based on the historical translation sequence. The second is to use the current decoding state as the query vector, perform cross-attention calculation on the cross-modal alignment representation, and obtain the multi-modal context most relevant to the current generation target. Thirdly, the probability prediction and translation output of the target word are completed after fusing the current decoding state and the multi-modal context. Since the semantic focus corresponding to each step of the decoding stage is not the same, the attention mechanism actually assumes the dual role of dynamic information routing and multi-modal evidence screening.

At the TTH time step, let the sequence of historical translations be:

$$Y_{<t} = \{y_1, y_2, \ldots, y_{t-1}\} \tag{21}$$

Here, yi denotes the ith target word lemma that has been generated. In order to ensure the autoregressive characteristics of the translation generation, the model first maps the historical target word units into embedding vectors, which are combined with position encoding to form the input representation at the decoder. Based on the masked self-attention mechanism, the decoded query state at the current time step can be expressed as follows:

$$Q_t = \text{MaskedSelfAttention}(Y_{<t}) \tag{22}$$

The function of Equation (22) is to extract the target-side context semantics of the current time step without accessing the information of future lemmas. Different from the self-attention in the source language encoding stage, the masking mechanism here ensures that the model can only rely on the already generated translation fragments for state updates, so as to satisfy the causal constraints of the sequence generation task.

After obtaining the decoded query state, the model further uses it as a semantic query vector for the current time step to guide the access to the cross-modal aligned representation Z. Here, Z refers to the unified multimodal representation obtained after cross-modal alignment and gated fusion as described above. To highlight the correlation between the current generation target and multi-modal semantics, this paper introduces multi-modal cross-attention calculation at the decoder, and its weight is defined as follows:

$$\alpha_t = \text{softmax}\left(\frac{(Q_t W_q)(ZW_k)^{\top}}{\sqrt{d}}\right) \tag{23}$$

Here, Wq and Wk represent the query mapping matrix and the key mapping matrix, respectively, and αt represents the attention strength of the current decoding state to each semantic unit in the multimodal alignment representation at the TTH time step. Equation (23) shows that the decoder does not utilize all cross-modal information equally, but assigns differentiated weights to different features according to the current generated semantics, thus achieving finer context retrieval.

Based on the attention weights, the multimodal context vector corresponding to the current time step can be written as follows:

$$C_t = \alpha_t(ZW_v) \tag{24}$$

Here, Wv is the value mapping matrix and Ct represents the most relevant multi-modal semantic summary under the current generation condition. The vector is essentially a weighted aggregation of the cross-modal alignment results by the attention mechanism, which can extract the information most relevant to the current lemma generation. The multimodal attention injection process is shown in Figure 4.

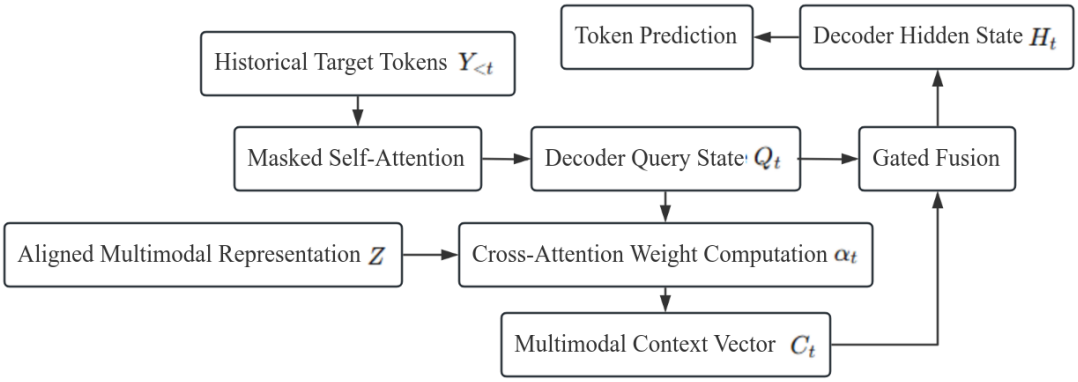

Figure 4: Multimodal Attention Injection Structure in the Translation Decoder

After obtaining the decoded query state Qt and the multi-modal context vector Ct at the current time step, the model does not directly use Ct for word prediction, but further adaptively integrates the historical semantics of the target side with the current cross-modal evidence through the gated fusion mechanism. The fusion gate value is defined as follows:

$$\beta_t = \sigma(W_f[Q_t; C_t] + b_f) \tag{25}$$

Here, [Qt;Ct] represents the vector concatenation operation, Wf and bf are the gating parameter matrix and bias term, respectively, and σ(·) is the Sigmoid activation function. Based on the gated weights, the fused hidden state at the current time step can be expressed as follows:

$$H_t = \beta_t \odot Q_t + (1 - \beta_t) \odot C_t \tag{26}$$

The meaning of Equation (26) is that when the current word generation depends more on the target side history context, the model will increase the contribution ratio of Qt. When the current lemma generation is more dependent on the multi-modal evidence after the image-text alignment, the model will increase the contribution ratio of Ct. In this way, the attention mechanism is no longer a simple weight assignment tool, but a dynamic control mechanism in translation generation combined with gating.

After multi-modal context injection and state fusion, the model makes a probability prediction for the target word at the current time step. The term generation probability is defined as follows:

$$P(y_t|Y_{<t}, Z) = \text{softmax}(W_o H_t + b_o) \tag{27}$$

Here, Wo and bo are the output layer weights and bias terms, respectively. Equation (27) shows that the prediction of the current lemma is jointly driven by two parts: one part comes from the target-side generation state formed by the historical translation sequence, and the other part comes from the contextual evidence provided by the cross-modal alignment results. Compared with the generation method that only relies on a single text context, this attention-driven production mechanism can better deal with translation samples with strong scene dependence, more semantic ambiguity and complex entity relationships.

From the generative logic, the attention mechanism in this paper has three prominent technical features. First, the decoded query state Qt is generated by masked self-attention, which can preserve target-side context information while following autoregressive constraints.

Secondly, the cross-attention mechanism enables the current time step to dynamically access the cross-modal alignment representation Z, avoiding the information compression problem caused by the fixed context vector. Thirdly, the gated fusion mechanism introduces a learnable trade-off between Qt and Ct, so that the model can adjust the proportion of language context and multimodal evidence use according to different generation stages. In general, this method extends the attention mechanism from a simple local weighting to a dynamic semantic scheduling mechanism for the whole process of English translation production.

From the perspective of computer technology implementation, the attention-driven translation generation process proposed in this paper has strong scalability. To further improve the expression ability of the model, multi-head cross-attention can be introduced into the decoder to model cross-modal associations in different semantic subspaces in parallel. Coverage mechanism or consistency constraint can also be combined to alleviate the problem of repeated generation and information omission in long sentence translation. However, under the framework of this paper, in order to maintain the interpretability of the model structure and the simplicity of engineering implementation, the decoding stage mainly retains three key links: mask self-attention, cross-modal cross-attention and gated fusion, so as to achieve a balance between translation quality and model complexity.

3 Data processing and experimental design

3.1 Multimodal English translation data construction and preprocessing

In order to verify the effectiveness of the proposed model in multimodal English translation production tasks, this paper constructs a multimodal English translation dataset containing text, image and translation annotations. The data source consists of three parts. The first part is English sentences and corresponding images in the public image corpus. The second is the manually supplemented Chinese-English parallel translation samples; Thirdly, a high consistency sample set is formed after secondary screening of low-quality descriptions, duplicate images and semantic inconsistent data in the original sample. In the process of data construction, the triple of "text-image-target translation" is taken as the basic unit, which requires that the source text can accurately describe the main scene of the image, the target translation is consistent with the semantics of the source language, and the image content can provide supplementary constraints on ambiguous words, entity relations or action semantics.

The preprocessing stage is divided into three steps: text processing, image processing and sample alignment. At the text end, a joint strategy of word segmentation and subword segmentation was used to convert the English sentence into a word element sequence. The samples with a length of more than 40 were truncated, and the insufficient part was filled. At the same time, the abnormal sentences with more than 30% special symbols are removed, and the case and punctuation formats are unified. At the image end, the original image is scaled to 224×224 resolution, and the visual block representation is generated by the method of combining regional division and feature extraction, and the images with low resolution, subject blur and content distortion are eliminated. The sample alignment stage focuses on checking the semantic consistency of the image and text, deleting the samples with inconsistent text and image topics, missing translations or mixed multiple translations, and dividing the training set, validation set and test set according to the ratio of 8:1:1. In order to reduce the training instability caused by the class distribution shift, we further apply statistical constraints on high-frequency words, low-frequency words and scene labels to keep the training samples relatively balanced in the distribution of semantic topics and sentence length. The statistical results of the processed dataset are shown in Table 2.

Table 2: Results of multimodal English translation dataset construction and preprocessing

Indicator	Value
Number of original image-text samples	18,500
Number of samples after initial screening	16,240
Number of final valid samples	15,600
Number of training samples	12,480
Number of validation samples	1,560
Number of test samples	1,560
Average English sentence length	18.7 tokens
Average Chinese sentence length	20.3 tokens
Maximum text length	40 tokens
Unified image size	224 × 224
Number of visual regions/patches	49
English vocabulary size	21,300
Chinese vocabulary size	24,100
Proportion of removed low-quality samples	15.68%

After the above processing, the dataset has good standardization in text length, image quality and image-text correspondence, which can provide a more stable data basis for subsequent cross-modal alignment training, attention injection decoding and model performance evaluation.

3.2 Experimental environment and parameter setting

In order to ensure the reproducibility of experimental results and the fairness of model comparison, this paper completed the model training and testing in a unified software and hardware environment. The experimental platform is configured with NVIDIA RTX 4090 GPU, Intel Core i9 processor and 64 GB memory, and the operating system is Ubuntu 22.04. The deep learning framework is PyTorch 2.1 with CUDA version 12.1. At the text encoder, the dimension of word vector was set to 256, the dimension of hidden layer was set to 512, and the number of multi-head attention heads was set to 8. At the visual encoder, the pre-trained visual backbone network was used to extract regional features and map them to a 512-dimensional shared semantic space. The number of encoder and decoder layers are both set to 6, the dimension of feedforward network is set to 2048, and the Dropout ratio is set to 0.1 to maintain a balance between model expressiveness and training stability. The Adam optimizer was used in the training phase, the initial learning rate was set to 1×10^{-4}, and the warm-up and cosine annealing strategies were combined for dynamic adjustment. The batch size was set to 32, the maximum training rounds were set to 50, and the training was stopped early when the validation set indicators were not improved for 5 consecutive rounds. The loss function is composed of translation cross-entropy loss and cross-modal alignment loss, in which the alignment loss weight is set to 0.3, and the gradient clipping threshold is set to 1.0 to suppress the gradient shock and explosion risk in the training process. In the experimental evaluation stage, the BLEU, METEOR and ROUGE-L indicators are used to measure the quality of translation. At the same time, the model parameter scale, single round training time and average inference delay are calculated to comprehensively test the actual performance of the constructed model in terms of translation accuracy, convergence efficiency and computational overhead.

3.3 Evaluation index and comparison scheme

In order to comprehensively evaluate the effectiveness of the proposed model in multimodal English translation production tasks, this paper sets up evaluation indicators from three dimensions of translation quality, semantic consistency and computational performance, and constructs a hierarchical comparison scheme consisting of text single-modal baseline, multi-modal enhanced baseline and the proposed model. Translation quality evaluation mainly uses three indicators: BLEU, METEOR and ROUGE-L. BLEU measures the matching degree between the generated translation and the reference translation at the n-gram level, and METEOR pays more attention to inflection, synonymous substitution and semantic coverage. ROUGE-L is used to test the performance of the translation in sequence structure and long-distance semantic preservation. Considering the emphasis on cross-modal semantic alignment, we further calculate the accuracy of image-text consistency and term disambiguation, which are used to evaluate the actual contribution of visual information to ambiguous word translation, entity relationship recognition and scene semantic constraints. At the same time, in order to reflect the engineering usability of the model, the parameter scale, single round training time, average inference delay and video memory occupancy are synchronously recorded in the experiment, so as to investigate the balance between performance improvement and computational overhead of the model. In terms of comparison schemes, this paper sets the Transformer text translation model as the basic baseline to test the gain before and after the introduction of visual modalities. A multi-modal translation model with only visual feature stitching is set to evaluate the effect of simple modal fusion strategies. The model containing only the cross-modal alignment module and the model containing only the attention enhancement generation module were set to distinguish the independent effects of the alignment mechanism and the attention mechanism. At the same time, the complete model of this paper is taken as the final experimental object, and its comprehensive advantages are analyzed through the horizontal comparison with the above models. In order to ensure a fair comparison, all the comparison models are run under the same data set, the same training rounds, the same optimizer and the uniform parameter scale constraints, and the average result is taken as the final evaluation basis after repeating the experiment three times on the test set.

Table 3: Evaluation indicators and comparison scheme Settings

Category	Specific Content	Description of Role
Translation Quality Metrics	BLEU, METEOR, ROUGE-L	Evaluate translation accuracy, semantic coverage, and sequence preservation capability
Semantic Consistency Metrics	Image-text consistency accuracy, term disambiguation accuracy	Evaluate the contribution of cross-modal information to scene understanding and ambiguity resolution
Computational Performance Metrics	Parameter size, training time, inference latency, GPU memory usage	Evaluate engineering efficiency and deployment feasibility of the model
Baseline Model 1	Transformer	Text-only unimodal baseline
Baseline Model 2	Visual Feature Concatenation Model	Used to examine the effectiveness of a simple visual fusion strategy
Baseline Model 3	Cross-Modal Alignment Only Model	Used to examine the independent contribution of the cross-modal alignment mechanism
Baseline Model 4	Attention-Enhanced Decoder Only Model	Used to examine the independent contribution of the attention-enhanced generation mechanism
Final Model	Proposed Model	Used to comprehensively evaluate the performance advantages of the complete framework

4 Experimental results and analysis

4.1 Performance Analysis of Multimodal English Translation Production Models

In order to test the comprehensive performance of the constructed model in the multimodal English translation production task, this paper compares it with the Transformer, the visual feature stitching model, the model with only cross-modal alignment module, and the model with only attention-enhanced decoding module. The experimental results are shown in Figure 5. The proposed model achieves the best performance in the main evaluation indicators, with BLEU, METEOR and ROUGE-L reaching 37.4, 32.5 and 41.3 respectively, which is 5.6, 4.4 and 5.9% higher than the basic Transformer model. Compared with the simple visual stitching model, the improvements are 4.2, 3.5 and 4.6% respectively. The results show that the direct stitching based on visual features can bring some gains, but it is difficult to fully explore the deep semantic relationship between images and texts. In contrast, the cross-modal alignment mechanism can significantly improve the matching quality between text and visual information, while the attention-enhanced decoding mechanism further improves the ability to select key contexts in the target word generation stage. On the whole, the proposed model is superior to various comparison models in translation accuracy, semantic coverage and sequence preservation ability, especially in the samples with strong scene dependence and complex entity relationships.

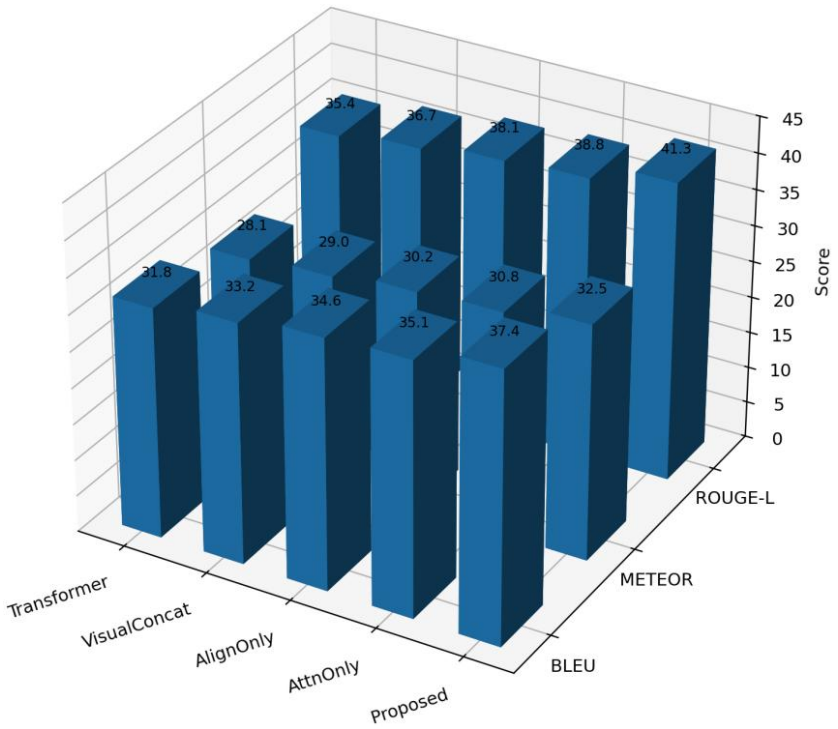

Figure 5: Multimodal English translation production model performance analysis

4.2 Analysis of Cross-modal alignment Effect

In order to further test the effectiveness of the proposed model in image-text semantic matching and cross-modal information collaboration, this paper analyzes the cross-modal alignment effect from three dimensions: image-text consistency accuracy, ambiguity resolution accuracy and entity alignment accuracy. The experimental results are shown in

Figure 6. The proposed model achieves 85.9%, 84.2% and 85.1% in the above three indicators respectively, which are significantly better than other comparison models. Compared with the basic Transformer, the three indicators are increased by 13.5, 15.3 and 14.8% respectively. Compared with the model using only visual feature concatenation, the proposed method improves 9.8, 11.7 and 11.1%, respectively. This shows that although simple modal stitching can introduce additional visual information, due to the lack of explicit semantic mapping and fine-grained matching mechanism, the image content is difficult to stably constrain the generation of translation. The model with only cross-modal alignment module has reached 81.7%, 79.8% and 80.6% on the three indicators respectively, indicating that the alignment mechanism itself has a direct effect on enhancing image-text matching, improving term disambiguation and strengthening entity recognition. Although the model with only attention-enhanced decoding module has certain advantages in the generation process of the target side, its overall effect is still lower than that of the full model due to the lack of front-end semantic alignment support.

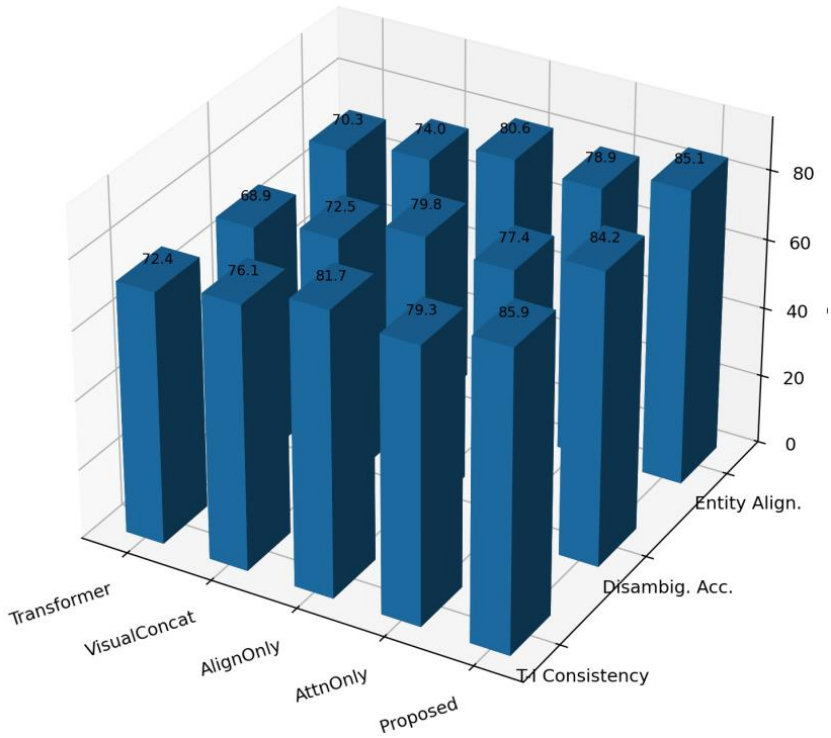

Figure 6: Analysis of cross-modal alignment effect

4.3 Analysis of the role of attention Mechanism

In order to further analyze the specific role of attention mechanism in multimodal English translation production, this paper starts from the translation generation performance under different sentence length conditions, and compares the attention-free decoder, attention-enhanced decoder and the full model of this paper. The experimental results are shown in Figure 7. As the sentence length increases, the BLEU values of each model show a downward trend, but the performance decline is significantly slowed down after the introduction of the attention mechanism. When the sentence length of the non-attention decoder increases from 10 to 40, the BLEU decreases from 35.2 to 28.7, a decrease of 6.5%. Attention-enhanced decoder decreased from 36.4 to 31.0, a decrease of 5.4%; The complete model in this paper decreases from 37.1 to 33.1, only decreasing by 4.0%. The results show that the attention mechanism can dynamically filter the most relevant semantic information in

the target word generation process, enhance the ability to maintain the key context in long sentence translation, and reduce the semantic omission caused by information compression. Especially in the samples with long sentences and relying on cross-modal semantic complement, the attention mechanism has a more obvious effect on improving the quality of translation.

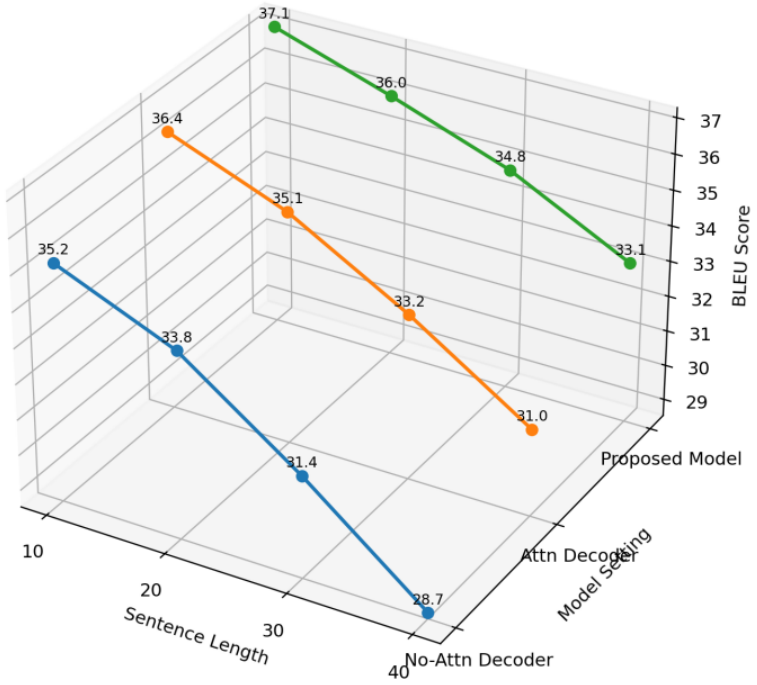

Figure 7: Analysis of the effect of attention mechanism

4.4 Application effect and limitation of the model are discussed

The experimental results show that the proposed model has a good application effect in multimodal English translation production tasks. On the one hand, the cross-modal alignment mechanism can map text information and visual information into a unified semantic space, which effectively alleviates the problems of insufficient ambiguity resolution, lack of scene constraints and unstable entity relationship recognition in traditional text translation. On the other hand, the attention-enhanced decoding mechanism enables the model to dynamically filter the context information most relevant to the current time step during the translation generation process, thereby improving the accuracy, semantic integrity and sequence coherence of the translation. Combined with the above experimental results, the proposed model is better than the comparison models in terms of BLEU, METEOR, ROUGE-L and image-text consistency accuracy, indicating that the method has certain application potential in intelligent translation assistance, image-text collaborative content generation and cross-lingual information service.

However, this model still has some limitations. First, the performance of the model largely depends on the pairing quality of the image-text samples. When the image information is fuzzy, the scene is incomplete, or the text description itself has deviation, the cross-modal alignment effect will be affected. Secondly, multi-modal feature encoding, alignment calculation and attention injection will increase the scale of model parameters and computational overhead, which is higher than the single-modal translation model in terms of training time, video memory occupation and inference delay. Third, although the current methods can improve the semantic collaboration ability of image and text, they still have

shortcomings in the processing of complex abstract semantics, cultural metaphor expression and deep pragmatic information. Therefore, future research needs to be further optimized in the aspects of high-quality data construction, lightweight model design, and complex semantic reasoning enhancement, so as to improve the adaptability and deployability of the model in real translation production scenarios.

5 Conclusion and Prospect

This paper focuses on the multimodal English translation production model combining cross-modal alignment and attention mechanism, constructs a unified technical framework of "multimodal input representation, cross-modal semantic alignment, and attention-driven decoding generation", and verifies the effectiveness of the model through experiments. The results show that the BLEU, METEOR and ROUGE-L of the proposed model on the test set reach 37.4, 32.5 and 41.3 respectively, which are 5.6, 4.4 and 5.9% higher than that of the basic Transformer model. The accuracy of image-text consistency, ambiguity resolution and entity alignment reaches 85.9%, 84.2% and 85.1%, respectively, indicating that the cross-modal alignment mechanism can effectively reduce the representation deviation between text semantics and visual semantics, and the attention mechanism can enhance the dynamic screening ability of key contexts in the process of translation generation. From the application effect, the model performs more stable in translation samples with strong scene dependence, complex entity relationship and obvious semantic ambiguity, indicating that it has certain application potential in intelligent translation assistance, image-text collaborative generation and cross-language information service scenarios. At the same time, there are still some shortcomings, mainly reflected in the model's strong dependence on high-quality paired samples of images and texts, and the quality of translation will be affected when the input image is blurred or the text description is incomplete. In addition, multi-modal coding, alignment calculation and attention injection increase the parameter scale and computational overhead, and the training efficiency and deployment cost still need to be further optimized. In the future, further progress can be made from expanding the scale of multi-domain and multi-scenario data, introducing lightweight cross-modal modeling methods, strengthening the ability of complex semantic and pragmatic reasoning, and improving the interpretability of the model, so as to enhance the adaptability, stability and engineering implementation ability of the model in the real English translation production environment.

About the Author

Shufang Wang was born in Zhengzhou, Henan. P.R. China, in 1981. She received the Doctoral degree from the University of Visayas, Cebu. Philippines. Now, she works in Zhengzhou Shengda University, foreign language school. Her research interest include computational intelligence, English teaching and literature.
E-mail: shirleywang202604@163.com

References

[1] Guo J, Su R, Ye J. Multi-grained visual pivot-guided multi-modal neural machine translation with text-aware cross-modal contrastive disentangling[J]. Neural Networks, 2024, 178: 106403.

[2] Guo J, Hou Z, Xian Y, et al. Progressive modality-complement aggregative multitransformer for domain multi-modal neural machine translation[J]. Pattern Recognition, 2024, 149: 110294.

[3] Liu Y, Liu D, Zhu S. Bilingual–visual consistency for multimodal neural machine translation[J]. Mathematics, 2024, 12(15): 2361.

[4] Shi X, Yang X, Cheng P, et al. Enhancing multimodal translation: Achieving consistency among visual information, source language and target language[J]. Neurocomputing, 2025, 620: 129269.

[5] Luo H, Guo Z, Wu Z, et al. Transformer-based vision-language alignment for robot navigation and question answering[J]. Information Fusion, 2024, 108: 102351.

[6] Zhao Y, Zhang Y, Sui X, et al. Me3a: A multimodal entity entailment framework for multimodal entity alignment[J]. Information Processing & Management, 2025, 62(1): 103951.

[7] Yao H, Wang L, Cai C, et al. Language conditioned multi-scale visual attention networks for visual grounding[J]. Image and Vision Computing, 2024, 150: 105242.

[8] Wei X, Kurtz C, Cloppet F. Enhancing vision–language contrastive representation learning using domain knowledge[J]. Computer Vision and Image Understanding, 2025, 259: 104403.

[9] Fang Z, Zou Y, Lan S, et al. Scalable multi-modal representation learning networks[J]. Artificial Intelligence Review, 2025, 58(7): 209.

[10] Liu Q, Hu J, Xiao Y, et al. Multimodal recommender systems: A survey[J]. ACM Computing Surveys, 2024, 57(2): 1-17.

[11] Zhang S, Liu J, Jiao Y, et al. A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(10): 1-22.

[12] Jin Y, Li J, Gu T, et al. Efficient multimodal large language models: A survey[J]. Visual Intelligence, 2025, 3(1): 27.

[13] Pu M, Luo B, Zhang C, et al. Text-vision relationship alignment for referring image segmentation[J]. Neural Processing Letters, 2024, 56(2): 64.

[14] Zhang T, Song B, Zhang Z, et al. Multimodal sentiment analysis based on multi-stage graph fusion networks under random missing modality conditions[J]. IET Image Processing, 2025, 19(1): e13310.

[15] Jiménez-Guarneros M, Fuentes-Pineda G. Multi-modal supervised domain adaptation with a multi-level alignment strategy and consistent decision boundaries for cross-subject emotion recognition from EEG and eye movement signals[J]. Knowledge-Based Systems, 2025, 315: 113238.

[16] Wang J, Xie H, Zhang S, et al. Multimodal fusion framework based on knowledge

graph for personalized recommendation[J]. Expert Systems with Applications, 2025, 268: 126308.

[17] Guo K, Tian D, Hu Y, et al. CFMMC-align: Coarse-fine multi-modal contrastive alignment network for traffic event video question answering[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10538-10550.

[18] Lin C, Cheng H, Rao Q, et al. M3SA: Multimodal sentiment analysis based on multi-scale feature extraction and multi-task learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 1416-1429.

[19] Liu S, Mao X, Zhao S, et al. Mer-clip: Au-guided vision-language alignment for micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2025.

[20] Schulze Buschoff L M, Akata E, Bethge M, et al. Visual cognition in multimodal large language models[J]. Nature Machine Intelligence, 2025, 7(1): 96-106.