Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061145

ARTICLE

Multimodal Neural Machine Translation Based on Knowledge Distillation and Anti-Noise Interaction

Erlin Tian[1], **Zengchao Zhu**[2,*], **Fangmei Liu**[2] **and Zuhe Li**[2]

[1]School of Software, Zhengzhou University of Light Industry, Zhengzhou, 450001, China
[2]School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, 450001, China
*Corresponding Author: Zengchao Zhu. Email: 332207050700@zzuli.edu.cn

Received: 18 November 2024; Accepted: 22 January 2025; Published: 16 April 2025

ABSTRACT: Within the realm of multimodal neural machine translation (MNMT), addressing the challenge of seamlessly integrating textual data with corresponding image data to enhance translation accuracy has become a pressing issue. We saw that discrepancies between textual content and associated images can lead to visual noise, potentially diverting the model's focus away from the textual data and so affecting the translation's comprehensive effectiveness. To solve this visual noise problem, we propose an innovative KDNR-MNMT model. The model combines the knowledge distillation technique with an anti-noise interaction mechanism, which makes full use of the synthesized graphic knowledge and local image interaction masks, aiming to extract more effective visual features. Meanwhile, the KDNR-MNMT model adopts a multimodal adaptive gating fusion strategy to enhance the constructive interaction of different modal information. By integrating a perceptual attention mechanism, which uses cross-modal interaction cues within the Transformer framework, our approach notably enhances the quality of machine translation outputs. To confirm the model's performance, we carried out extensive testing and assessment on the extensively utilized Multi30K dataset. The outcomes of our experiments prove substantial enhancements in our model's BLEU and METEOR scores, with respective increases of 0.78 and 0.99 points over prevailing methods. This accomplishment affirms the potency of our strategy for mitigating visual interference and heralds groundbreaking advancements within the multimodal NMT domain, further propelling the evolution of this scholarly pursuit.

KEYWORDS: Knowledge distillation; anti-noise interaction; mask occlusion; door control fusion

1 Introduction

In the realm of Natural Language Processing (NLP), the pursuit of machine translation stands as a cornerstone, dedicated to the automated translation of textual content across diverse languages [1]. Among the myriad of NLP endeavors, multimodal neural machine translation (MNMT) appears particularly significant, given its extensive utility in global communication, cultural exchanges, entertainment, and media design. With the rapid advancement of deep learning and neural network technologies, the development of efficient and correct machine translation systems has become an urgent research priority [2]. Our MNMT model advances the frontier of translation technology by incorporating a variety of data streams, such as images, videos, and audio. Empirical evidence from existing studies shows that the incorporation of added visual information can significantly boost the efficacy of neural machine translation. Therefore, we are actively exploring strategies to integrate multimodal data seamlessly, to enhance translation quality, a pursuit that has become a central focus in the contemporary research landscape.

In recent years, MNMT researchers have concentrated on crafting sophisticated multimodal fusion frameworks [3] to bridge the semantic gap between images and their accompanying texts. However, the challenge of visual noise has often been neglected. In practical scenarios, achieving a flawless alignment between image content and pure textual data is highly challenging. Moreover, images and textual information often show weak correlations, making visual noise a persistent concern [4]. The image information we derive from analyzing the text is not necessarily all that the image represents. Redundant visual features can make our recognition more difficult. As shown in Fig. 1, the overall message of the human-observed image description is "brown dog is running after the black dog on the pebbles by the beach," but the text provides only the content message "brown dog is running after the black dog". The resulting German translation is "Der braune Hund rennt dem schwarzen Hund hinterher". In this process, the focus of the translation should be more on the "running" behavior of the "brown dog" and the "black dog" rather than on the environmental message "pebbles, moss, and beach". This extra environmental information is not what machine translation is looking for. Redundant image information may lead to confusion of the translation focus, which in turn affects the accuracy of the translation.

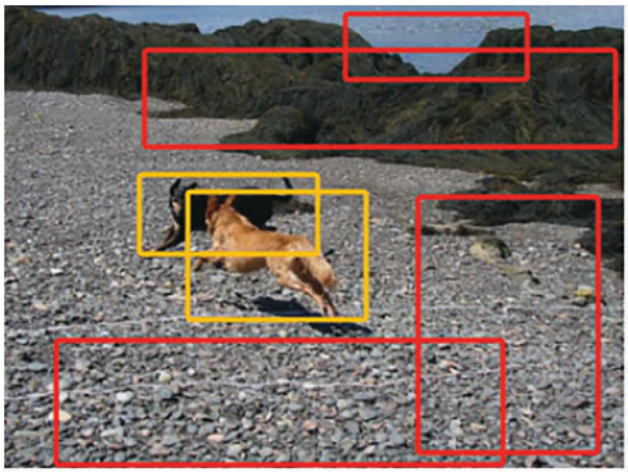

Figure 1: An example of "brown dog is running after the black dog." It underscores the imperative to account for visual noise within the translation model. In this scenario, the yellow color is the core content of the text description, while the red color symbolizes the visual interference noise

In the realm of Multimodal Neural Machine Translation (MNMT), a critical challenge enhancing the robustness of text-image integration. Researchers emphasize the critical need to find and filter out discordant visual-textual information, a pivotal step for improving translation accuracy. The accuracy of machine translation is significantly enhanced by a well-designed multimodal fusion strategy. To address this, researchers have introduced a series of innovative approaches: (1) The integration of a multimodal attention mechanism [5]. This mechanism enhances the effective integration of visual and textual features through cross-modal attention and adaptive feature selection; (2) The adoption of a multimodal Transformer fusion method [6]. By using the Transformer architecture to independently encode textual and visual features, it eases their integration via a multi-modal cross-modal attention mechanism; (3) The development of a gated fusion technique [7]. This technique ensures the alignment of textual semantic representations with their visual equivalents, ensuring coherent fusion that propels the progress of MNMT.

While current models have indeed advanced the robustness [8] of MNMT systems, particularly in their handling of noisy data, these efforts have centered on using visual information to refine conventional

machine translation processes [9]. However, they have not considered the influence of visual noise within the multimodal feature fusion architecture. Considering this oversight, our study introduces a pioneering multimodal interaction fusion strategy that is anchored in the Transformer architecture [10], specifically designed to combat the interference posed by visual noise. This strategy incorporates a bidirectional knowledge distillation technique [11] to forge a foundational correlation between textual and visual data within the visual Transformer encoder. Concurrently, we have implemented a masking occlusion mechanism to develop an attention module adept at discerning graphic-image relationships, thereby enhancing the extraction of pertinent visual features through the masking process. Furthermore, we have developed a cross-modal gating mechanism to mitigate noise, ensuring the effective integration of multimodal features within a sequence-to-sequence (seq2seq) framework [12]. This approach enhances the efficient fusion of these features.

In contrast to traditional studies [13], our proposed approach for image-text fusion significantly outperforms existing models in performance metrics. This method eases the integration of visual elements that align closely with the image attributes. The KDNR-MNMT model proposed in this study is an innovative improvement of the traditional multimodal neural machine translation framework, which consists of two core components. The first part, shown on the right side of Fig. 2, is based on the traditional neural machine translation framework and includes multiple encoding layers, decoding layers, and positional encoding. The second part, shown on the left side of Fig. 2, consists of three modules with different functions, namely, the bidirectional knowledge distillation module, the masking module, and the gating fusion module, which work together to enhance the model performance. The result robustly proves the potency of our multimodal engagement tactic for translating tasks. A thorough analysis of our experimental outcomes not only corroborates that our model eclipses existing ultramodern methods within the MNMT domain, but also underscores its substantial enhancement of machine translation performance. Moreover, we place a premium on the interpretability of our model. Through a meticulous examination of the experimental data, we further confirm the generalizability of our proposed technique, ensuring the model's transparency and reliability. The principal contributions of this research are three-pronged:

(1) We introduce a pre-training model designed to capture multimodal graphic features by integrating global visual and textual features through a process of bidirectional knowledge distillation. Our findings show that these synthesized multimodal graphic features are pivotal for supporting training stability.

(2) We present an innovative noise-robust multimodal fusion method that uses mask-obscured modal relations. This technique integrates synthetic graphical elements with visually masked features, empowering our model to extract pertinent details more efficiently and lessen the effects of visual noise.

(3) We have developed a gating system that employs a cross-modal interaction masking mechanism, tailored for the representation and fusion of noise-resistant multimodal features within noisy contexts. This system forms the backbone of our KDNR-MNMT model.

The next parts of this research are organized in the following manner: Section 2 offers an exhaustive examination of pertinent literature concerning MNMT. In Section 3, we outline the overarching structure of the suggested model, complemented by an in-depth description of its components. Section 4 underscores the efficacy of the model and conducts an in-depth assessment of each part through a series of meticulous experiments. Concluding with Section 5, which presents a comprehensive overview of the research findings.

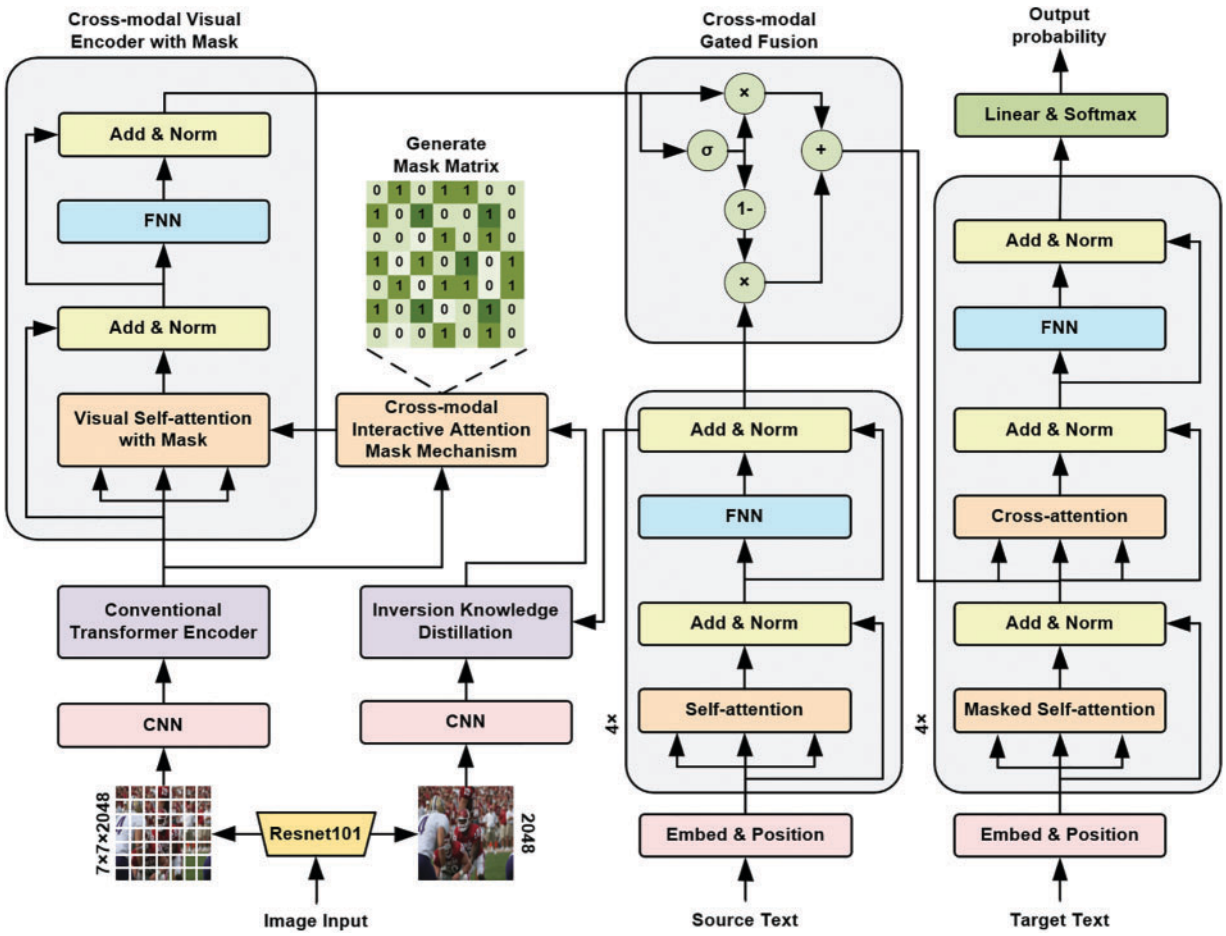

Figure 2: Overall flowchart of the KDNR-MNMT model

2 Related Work

Here, we start with a succinct summary of the prevailing scholarly progress on the MNMT platform as presented in Section 2.1. Subsequently, in Section 2.2, we elaborate on knowledge distillation techniques to synthesize information from different modalities. Moving forward to Section 2.3, we delve into the masking occlusion strategy we have adopted, providing a theoretical validation of its reliability. To conclude this section, we elucidate the gating mechanism we have developed in Section 2.4, highlighting its effectiveness in the field of cross-modal fusion.

2.1 Multimodal Neural Machine Translation

Multimodal Neural Machine Translation (MNMT) fuses visual and textual data to improve the accuracy and fluency of translation. The technique processes the source language text and the corresponding images simultaneously through an encoder, which accurately captures key visual features with the help of an attention mechanism and deeply fuses them with textual features. Subsequently, the decoder takes this fused multimodal information into account when generating the target language text. This approach can effectively capture more comprehensive and rich contextual information, thus enhancing the accuracy and naturalness of the translation results. At the onset of machine translation inquiry, researchers predominantly

concentrated on encoder-decoder frameworks predicated on recurrent neural networks, as supported by seminal investigations [14–16]. Pioneers in this field, such as Zhao et al. [17], have integrated visual region features with textual data by using object detection features and region-related attention mechanisms. Nishihara et al. [6] presented a supervised cross-modal attention mechanism to harmonize textual and pictorial elements. Song et al. [18] integrated a co-attentional graphic refresh module across the Transformer encoder's layers for the alignment of multimodal attributes. Additionally, Yao et al. [10] employed the multimodal Transformer to harmonize visual and textual features. Yin et al. [4] introduced a graph-oriented method for multimodal neural machine translation, enabling the extraction of multifaceted features via a synchronized text-image attention system. Lin et al. [19] employed a selective filtering approach coupled with a contextually adaptive capsule network to synthesize visual features. Despite these innovations, the task stays challenging and resource-intensive due to constraints in the quantity and quality of labeled images.

Our study addresses this by aiming to perform machine translation with prior access to images, thereby transcending data limitations. In summary of the challenges found, we propose a pioneering approach aimed at boosting translation accuracy by incorporating image data at an early stage of the translation workflow, thereby reducing the reliance on lots of high-quality annotated images. This approach not only bolsters translation accuracy but also mitigates the costs associated with data collection and processing.

2.2 Knowledge Distillation Strategy

Caruana et al. [20] and Hinton et al. [21] pioneered the concept of Knowledge Distillation, a strategy crafted to convey insights from an expansive and complex Teacher Model to a streamlined, high-performance Student Model. The core principle of this method is to empower the Student Model to match the Teacher Model's efficacy, all the while diminishing the computational and memory demands, which results in model streamlining and performance augmentation. Romero et al. [22] extended the applicability of knowledge distillation by conveying wisdom from layers within the model's intermediate representation. Yim et al. [23] introduced a technique for knowledge acquisition that uses the flow of information between layers, achieved by computing the interaction between the layer features. In the multimodal fusion domain, Gupta et al. [24] pioneered the transfer of supervised knowledge between images of varying modalities. In contrast, Yuan et al. [25] presented a symmetric distillation network tailored for text-to-image synthesis tasks.

Capitalizing on these pioneering investigations, we present an innovative knowledge distillation module within this study, designed to address the issues of limited data availability and the high expenses associated with annotation in the realm of multimodal machine translation. This research introduces our module, which seamlessly amalgamates diverse modalities and yields more optimized translation results within constrained data scenarios. By harnessing knowledge distillation techniques to craft multimodal attributes, our approach transcends the limitations of sparse datasets.

2.3 Masking Strategy

Masking strategies are pivotal in the realms of machine learning and deep learning, particularly for handling invalid or irrelevant input data and bolstering model generalization. These strategies have seen extensive application across various domains, including NLP, image processing, and multimodal learning. In the context of text-image fusion, masking strategies are instrumental in modulating the exchange of information between different modalities. Song et al. [26] introduced a novel pre-training task for language generation that employs a sequence-to-sequence masking technique. Li et al. [27] proved that by applying masks to certain modalities, the model is guided to focus more intently on and integrate information from other modalities, thus enhancing a comprehensive understanding of the context. Huang et al. [28] used a masking technique to obscure extraneous details, encouraging the model to concentrate on extracting

pertinent features from the contextual fabric, within a multilingual pre-training framework aimed at achieving cross-linguistic representations. This method bolsters the model's resilience and precision. The method first utilizes the self-attention mechanism to process the initial text features and then combines these features with image features to form graphic-text fusion features. Next, by evaluating the weight values and zeroing the weight values lower than the preset threshold, this achieves masking occlusion and effectively filters out the feature weights with low attention. Finally, the adjusted weight information is multiplied with the fusion feature matrix to reduce redundant information and optimize the feature representation to enhance the performance of multimodal neural machine translation.

The efficacy of masking strategies has been confirmed across a multitude of pre-training representation learning endeavors. Building on this foundation, our current study delves into the potential of masking strategies in addressing the challenge of noise-robust multimodal fusion in MNMT.

2.4 Cross-Modal Gating Mechanism

In this study, we introduce a pioneering gating mechanism that is instrumental in the realms of machine learning and deep learning, specifically within the domains of sequential data handling and attention-based modeling. This mechanism judiciously regulates the flow of information, deciding which feature details to preserve or discard during model execution. Within the Transformer framework, our gating mechanism skillfully orchestrates the allocation of attention, guiding the model to prioritize the pivotal elements of feature data. Raffel et al. [29] pioneered the use of the gating mechanism in a text-to-text framework, enhancing machine translation accuracy by limiting the integration of known information. Expanding on this concept, Bao et al. [30] applied a specialized gating mechanism known as pseudo-masking in pre-training tasks for language comprehension and generation, using it as a distinct gating technique. Consequently, we have integrated learnable parameters that empower the model to self-optimize the gating strategy throughout the training phase. This approach not only effectively mitigates overfitting but also maximizes performance.

The incorporation of our gating mechanism bestows the model with enhanced flexibility and selectivity in data handling, significantly boosting the overall efficacy and adaptability of the KDNR-MNMT model. This advancement positions our research at the forefront of leveraging gating mechanisms to navigate the complexities of multimodal neural machine translation.

3 Methodology

Here, we introduce our innovative approach for noise-resistant multimodal neural machine translation, which uses bidirectional knowledge distillation. This architecture incorporates the KDNR-MNMT framework within the Transformer model, encompassing four distinct sub-networks: (1) cross-modal feature encoder; (2) robust masked matrix image encoder; (3) cross-modal gated fusion module.

3.1 Cross-Modal Feature Encoder

Typically, we employ the conventional positional embedding layer to incorporate the input data. Let $x_j = \left\{x_1^j, \cdots, x_I^j\right\}$ and v_j denote the lengths of the source text and its accompanying image for a given data pair j, I representing the length of the source text x_j. Officially, the source sentence is represented by E_j^x, which includes both word and positional embeddings within a text embedding layer. Concurrently, the global visual and segmented features are represented by $E_j^{\mathrm{V}} \in \mathbb{R}^{I \times d_1}$ and $E_j^v \in \mathbb{R}^{m \times m \times d_2}$, which are derived from visual feature extraction layers utilizing ResNet-101 [31].

Initially, we employ straightforward average pooling to condense the array of word embeddings into comprehensive textual features, as detailed by Li et al. [27]. Subsequently, the comprehensive textual features are progressively introduced into a multimodal feature $\bar{t}$ generator, easing the computation and derivation of enriched multimodal features m. See Eqs. (1) and (2):

$$\bar{t} = \frac{1}{I}\sum_{j=1}^{I} E_j^x \tag{1}$$

$$m = \text{unpool}\left(W^t \bar{t}\right) \tag{2}$$

where the fully connected (FC) layer W^t maps the aggregated textual features $\bar{t}$ into the image domain. Subsequently, an averaging approach is applied to the lower-dimensional vectors, enabling the calculation of more comprehensive multimodal feature maps. The dimensionality $m \in \mathbb{R}^{P*2048}$ matches that of the final convolutional layer's output in the teacher model.

Within the multimodal encoding layer, we integrate multimodal features m with textual features $\bar{t}$ to reconstruct the multimodal features as queries. See Eq. (3):

$$\widetilde{x} = \left[t; mW^m\right] \in \mathbb{R}^{(I+P)*d} \tag{3}$$

where the P is the size of the multimodal feature. Looking at modal fusion from a graph-theoretic perspective, each source language token can be viewed as a node in the graph. So, each part of a multimodal feature can be viewed as a virtual token that is incorporated into the structure of the graph formed by the source tokens for inter-modal fusion. Concurrently, the key and value vectors are kept being textual features $\bar{t}$, with the computation of the multimodal encoder layer detailed later. See Eq. (4):

$$\mathrm{H}_{x_j}^l = \text{Multihead}\left(m, \mathrm{E}_j^x, \mathrm{E}_j^x\right) = \text{Concat}\left(head_j^1, \ldots, head_j^M\right) \tag{4}$$

where M stands for the count of attention heads, Multihead$(\cdot)$ signifies the Multi-headed Attention mechanism, and $l = (0, \cdots, 3)$ shows the index of the Transformer layer. Officially, the multi-head attention's resultant output is figured out in a later manner. See Eq. (5):

$$head_j^{c\in[1,M]} = \sum_{k=1}^{n} \alpha_{ik}\left(\mathrm{E}_{jk}^x \mathrm{W}_{j,c}^V\right) \tag{5}$$

where n corresponds to the length of x_j. Additionally, the weight coefficients α_{ik} are derived through the application of the softmax function. See Eq. (6):

$$\alpha_{ik} = \text{softmax}\left(\frac{\left(\widetilde{x}\mathrm{W}_{j,c}^Q\right)\left(\mathrm{E}_{j_k}^x \mathrm{W}_{j,c}^K\right)^{\mathrm{T}}}{\sqrt{d}}\right) \tag{6}$$

where α_{ik} stands for the attention matrix that integrates text and multimodal features through dot product attention. $W_{j,c}^V$, $W_{j,c}^Q$ and $W_{j,c}^K$ denote the respective parameter matrices. Ultimately, a position-wise feed-forward neural network is employed to refine the state of each position within the output sequence $F_{x_j}^l$. See Eq. (7):

$$\mathrm{F}_{x_j}^l = \text{FFN}\left(\mathrm{H}_{x_j}^l\right) \tag{7}$$

3.2 Masking Matrix Image Encoder

In the left-hand illustration of this section, we pay special attention to presenting the visual encoder with a masking mechanism. To minimize the model's parameter count, we opted for incorporating solely one layer of the Transformer architecture within our image encoder's design.

3.2.1 Legacy Transformer Encoder for Vision

Within the scope of this research, we use a pre-trained ResNet-101 model to decompose the spatial characteristics of the segmented image, resulting in 49 distinct spatial regions, each represented by a 7 × 7 × 2048-dimensional vector. Subsequently, these features were transformed by linear transformation into a 49 × d feature matrix, where d is the dimensionality of the word embeddings. To set up the internal correlation among the 49 image regions, we have used a conventional Transformer encoder to generate contextual representations for these local spatial region features. This approach enriches the semantic information available for multimodal feature fusion, denoted as H_{v_j}. See Eqs. (8) and (9):

$$\mathrm{H}_{v_j} = \mathrm{Multihead}\left(\mathrm{E}_j^v, \mathrm{E}_j^v, \mathrm{E}_j^v\right) \tag{8}$$

$$\mathrm{F}_{v_j} = \mathrm{FNN}\left(\mathrm{H}_{v_j}\right) \tag{9}$$

3.2.2 Cross-Modal Mask Masking Mechanisms for Vision

Motivated by the work of Li et al. [32], this section elaborates on the development of our cross-modal visual encoder equipped with a masking part. To filter out extraneous visual data during the multimodal integration phase, we crafted an interactive cross-modal attention masking mechanism, illustrated in Fig. 3. This approach initially eases the interaction between textual and visual attributes, then assesses the relationship between the 49 regional characteristics and the textual data. See Eqs. (10) and (11):

$$\mathrm{Matrix}_{v_j} = \mathrm{softmax}\left(\frac{\mathrm{F}_{v_j} \times \left(\mathrm{F}_{x_j}^l\right)^{\mathrm{T}}}{\sqrt{d}}\right) \tag{10}$$

$$\mathrm{Matrix}_{x_j} = \mathrm{softmax}\left(\frac{\mathrm{F}_{x_j} \times \left(\mathrm{F}_{v_j}^l\right)^{\mathrm{T}}}{\sqrt{d}}\right) \tag{11}$$

where $\mathrm{Matrix}_{v_j} \in \mathbb{R}^{49\times n}$ stands for the focus of the 49 regional features on each word of the corresponding source text, while $\mathrm{Matrix}_{v_j} \in \mathbb{R}^{49\times n}$ signifies the focus of each word in the source text on the 49 regions of the associated image. Subsequently, we calculated Matrix_{v_j} and Matrix_{x_j} through an interactive process as detailed below. See Eq. (12):

$$\mathrm{Mask}_j = \mathrm{Matrix}_{v_j} \times \mathrm{Matrix}_{x_j} \tag{12}$$

where Mask_j is a matrix that represents the correlation between 49 localized regions of an image and the source sentence. This matrix generates a mask matrix based on the informational importance of the local regions of the image and uses a predefined threshold $prob_r$ to determine which image regions need to be masked. See Eq. (13):

$$m_r = \begin{cases} 1, & prob_r \geq p, (r = (1, 2, \ldots, 49)) \\ 0, & prob_r < p \end{cases} \tag{13}$$

where p is a scalar hyperparameter implemented to mask features from less important visual regions. Our strategy is crafted to guarantee that each image prominently displays the visual region most pertinent to its corresponding source text. Subsequently, we construct a masked knowledge matrix, where the image region associated with $m = 0$ is set to False, and that with $m = 1$ is set to True. This binary masking technique allows us to isolate and emphasize the relevant visual content.

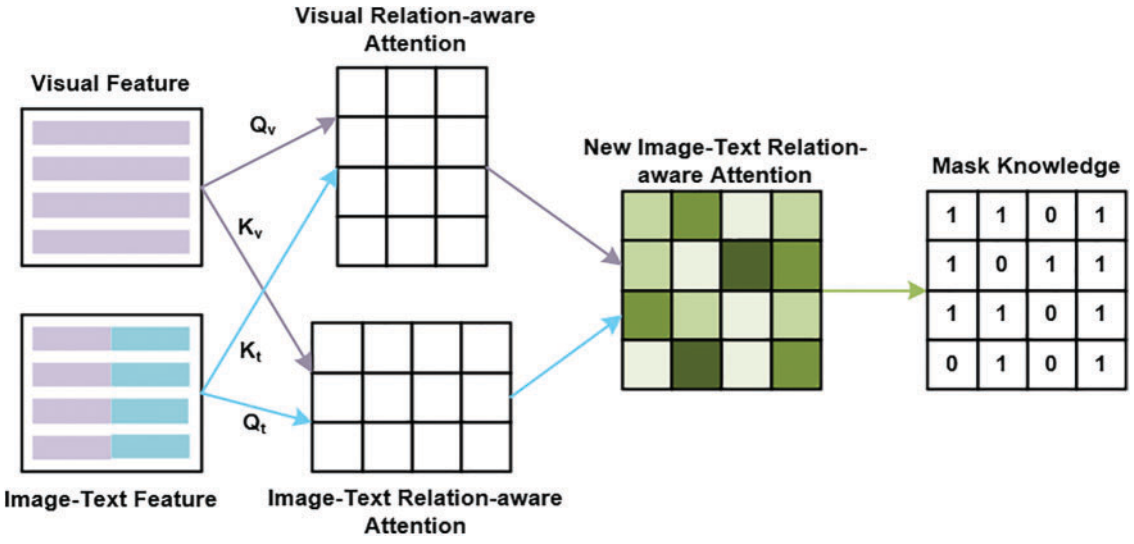

Figure 3: Cross-modal interaction attention masking mechanism modules

We deploy a cross-modal visual code equipped with these masks to accurately extract and integrate valid visual information, ensuring that our multimodal approach effectively captures the essence of both textual and visual data for enhanced translation tasks. See Eqs. (14) and (15):

$$\hat{\mathrm{H}}_{v_j} = \text{Multihead-mask}\left(\mathrm{F}_{v_j}, \mathrm{F}_{v_j}, \mathrm{F}_{v_j}\right) \tag{14}$$

$$\hat{\mathrm{F}} = \mathrm{FFN}\left(\hat{\mathrm{H}}_{v_j}\right) \tag{15}$$

In this context, Multihead-mask $(*)$ refers to self-attention that incorporates masking information, while Multihead-mask designed to drop weakly relevant visual details.

3.3 Cross-Modal Gating Fusion

In this section, we employ a cross-modal gating fusion technique to integrate textual features and extract relevant visual features. Given that in the previous feature fusion process, we have utilized a bidirectional knowledge distillation mechanism and a masking occlusion mechanism to achieve high-quality interaction between visual and textual features, we can appropriately reduce the weight of textual features in the final synthesis in the gating system. See Eqs. (16) and (17):

$$\Omega = \mathrm{Sigmoid}\left(\mathrm{W}_{\Omega}\hat{\mathrm{F}}_{v_j} + \mathrm{U}_{\Omega}\mathrm{F}_{x_j}\right) \tag{16}$$

$$\mathrm{H}_{g_j} = (1 - \Omega)\,\mathrm{F}_{x_j} + \Omega\hat{\mathrm{F}}_{v_j} \tag{17}$$

where W_{Ω} and U_{Ω} represent trainable parameters of the model. The final output, denoted as H_{g_j} is directly input into the target sentence decoder (shown on the right side of Fig. 2) for predicting the translation.

4 Experiments

4.1 Data

Dataset: In this research, we performed a comprehensive set of experiments using the Multi30K dataset [33], a renowned human-annotated resource within the multimodal machine translation (MMT) community. This collection associates each textual item with its matching JPG image from the Flickr30K dataset, providing human-translated versions of the texts across English, German, French, and Czech languages. The Multi30K dataset is designed with 29,000 samples distributed for training, 1014 for the validation set, and 1000 reserved for the Test2016 segment, for each language pair within the dataset. Additionally, for a comprehensive evaluation of our model, we performed assessments on both the Test 2017 and MSCOCO datasets, which each contribute an extra 1000 examples. Every textual entry in the Multi30K dataset is complemented by a corresponding JPG image from the Flickr30k collection [34].

Data Preprocessing: We standardized the preprocessing of both source and target utterances using the official Multi30K script, encompassing word list construction, application of byte-pair encoding (BPE), and up to 10,000 merge operations. This process prepared the sentence pairs for our experiments. For the visual part, we used a pre-trained ResNet-101 model to derive visual characteristics, resulting in a 2048-dimensional vector encapsulating global information and 49 local spatial region vectors, with each vector being $7 \times 7 \times 2048$ in size. These comprehensive features offer a wealth of visual data for our multimodal machine translation model.

Metrics: To comprehensively assess translation quality, we employ two main metrics: (1) the 4-gram BLEU score [35], which measures the precision and fluency of translations by analyzing the overlaps of 4-gram between the text produced by machines and the reference translations created by humans, thereby reflecting the overall excellence of the translations; (2) The METEOR score [36], which considers the precision and recall of translation segments, measuring the translation's alignment with reference texts to gauge its fluency and comprehensiveness. These metrics collectively offer a multifaceted assessment framework for translation quality, ensuring a balanced view of accuracy and readability.

Parameter Settings: The architecture we propose is built upon the Transformer framework and consists of just 4 layers each for the encoder and decoder stacks, ensuring that the model supports a low parameter count. We assign a dimension of 128 to the hidden states for both the encoder and decoder, while the inner layer of the feed-forward network is configured to have a dimension of 256. The learning rate is set up at 0.005, and we cap the maximum number of tokens at 4096. Additionally, the learning rate is adjusted over a warm-up phase of 2000 steps, with a label smoothing parameter configured to 0.1. We employ the Adam optimizer, using momentum parameters set to $\beta 1, \beta 2 = (0.9, 0.98)$. We incorporate 4 attention heads in this configuration and configure the dropout rate to 0.3 to prevent overfitting. The constraint dimension's width is configured at 5 units. Our KDNR-MNMT model undergoes training on an NVIDIA GTX 3090 GPU, utilizing fp16 precision.

4.2 Results on the EN→DE Translation Task

As detailed in Table 1, we offer a comparative examination of our KDNR-MNMT model in conjunction with existing state-of-the-art models for the task of translating from English to German (EN→DE). Our model outperforms its peers significantly when measured by the critical BLEU and METEOR benchmarks, essential indicators of translation excellence. Below, we will exhaustively summarize and compare the existing models in several key areas.

Table 1: Comparison results on Multi30k EN→DE task on BLEU ("B") and METEOR ("M") metrics

Model	Multi30K EN→DE					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
Fusion-conv [37]	37.0	57.0	29.8	51.2	25.1	46.0
VAG-NMT [38]	–	–	31.6	52.2	28.3	48.0
Del+Obj [39]	38.0	55.6	–	–	–	–
DCCN [22]	39.7	56.8	31.0	49.9	26.7	45.7
GMNMT [4]	39.8	57.9	32.2	51.9	28.7	47.6
OVC+L_v [40]	–	–	32.4	52.3	28.6	48.0
WRA-guided [17]	39.3	58.3	32.3	52.8	28.5	48.5
Our Transformer-Based Systems						
Transformers (NMT) [41]	40.96	58.35	32.59	51.21	29.16	48.37
Doubly-ATT [42]	41.44	59.08	33.15	52.34	29.22	48.41
Multimodal self-att [43]	41.50	58.52	32.51	51.33	29.10	48.48
Gated Fusion MNMT [44]	41.58	58.88	33.01	51.90	30.04	48.95
Mutual Information [45]	41.77	58.60	34.58	–	30.61	–
DLMulMix [46]	41.77	58.93	33.07	51.85	29.90	49.09
Our model	**42.36**	**59.87**	**35.01**	**54.42**	**31.07**	**49.67**

Comparison with Text-to-Text NMT: Our model outperforms the traditional Neural Machine Translation (NMT) baseline, with notably higher scores on the test set. This significant advancement not only underscores the superiority of our enhanced model but also confirms the effectiveness of our MNMT approach in using image data. By integrating a multimodal feature fusion strategy and an optimized model architecture, our model surpasses conventional methods in enhancing both the precision and contextual comprehension of machine translation. The KDNR-MNMT model excels in noise processing and filtering visual information, ensuring robust performance in diverse visual contexts.

Comparison with Existing MNMT Systems: Our proposed KDNR-MNMT model has proved a notable performance lead in our experimental series, proving its superiority over the current ultramodern SOTA models. In the domain of translating from English to German, our model has nearly realized a 1.0-point increase in both the BLEU and METEOR scores. This significant leap in performance is primarily due to our approach's ability to efficiently sift through the global image for key information while effectively sidelining irrelevant visual noise. By employing this nuanced information filtering mechanism, we ensure that the feature data presented to the decoder is both precise and valuable, thereby markedly elevating the quality of multimodal neural machine translation. Armed with this refined information processing strategy, the KDNR-MNMT model highlights its robust competence in navigating the intricacies of complex multimodal environments for effective translation tasks.

Overall, our proposed approach has achieved substantial enhancements across all critical evaluation metrics, outperforming the experimental outcomes of prior researchers and standing on par with the current ultramodern SOTA methods. These impressive results underscore the robustness of our method. By

integrating a bidirectional knowledge distillation mechanism, we adeptly distill pertinent visual information, and by tactfully masking less relevant visual data, we guide the model to concentrate on the most valuable content. Additionally, the gating mechanism we've integrated enhances the cross-modal feature fusion process, making the model more streamlined and correct in managing multimodal data. The constructive collaboration of these techniques not only bolsters the performance of the translation task but also paves new avenues and research directions within the multimodal machine translation domain.

4.3 Results on the EN→FR Translation Task

To validate the resilience to noise and the broader applicability of our model, additional tests were conducted on the Multi30K dataset for the translation task from English to French, as outlined in Table 2. The key takeaways from our analysis are as follows: firstly, our proposed model garners substantial improvements in the principal evaluation metrics over existing models, mirroring our findings in the English to German (EN→DE) task. Secondly, our MNMT model outperforms the baseline NMT model, which relies solely on textual data, by integrating image information, thereby confirming the model's adeptness at using visual cues to bolster translation quality. Ultimately, within the context of the robust NMT baseline for English-to-French translation, our strategy has significantly outperformed the prevailing state-of-the-art (SOTA) methodologies, demonstrating robust competitiveness. The results from the English-to-French translation endeavor confirm the effectiveness and wide-ranging applicability of our introduced KDNR-MNMT model.

Table 2: Comparison results on Multi30k EN→FR task on BLEU ("B") and METEOR ("M") metrics

Model	Multi30K EN→FR					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
Fusion-conv [37]	53.5	70.4	51.6	68.6	–	–
VAG-NMT [38]	–	–	53.8	70.3	45.0	64.7
Del+Obj [39]	59.8	74.4	–	–	–	–
DCCN [22]	61.2	76.4	54.3	70.3	45.4	65.0
GMNMT [4]	60.9	74.9	53.9	69.3	–	–
OVC+L_v [40]	–	–	54.2	70.5	45.2	64.6
WRA-guided [17]	61.8	76.3	54.1	70.6	43.4	63.8
Our Transformer-Based Systems						
Transformer (NMT) [41]	60.33	75.64	53.45	71.57	43.61	65.72
Doubly-ATT [42]	60.94	75.99	53.63	71.56	44.78	65.35
Multimodal self-att [43]	61.44	75.77	54.56	71.62	44.59	65.08
Gated Fusion MNMT [44]	61.24	76.26	54.15	71.77	44.29	64.91
DLMulMix [46]	62.23	76.85	55.18	73.37	44.42	66.41
Our model	**62.88**	**77.01**	**55.19**	**72.44**	**46.31**	**65.96**

4.4 Ablation Experiment

To confirm the performance of our suggested KDNR-MNMT model, we carried out an extensive series of ablation studies covering both the English-to-German and English-to-French translation tasks. This investigation encompassed an ablation analysis of the hyperparameter p, alongside a dissection of the various model components, providing insights into their contributions to the model's overall performance.

Ablation study of hyper-parameter *p*: To confirm the robustness against noise and the broader applicability of our model, we performed an array of ablation studies targeting the machine translation domain, placing particular emphasis on how the hyperparameter *p*'s threshold size impacts model efficacy. As displayed in Table 3, our analysis investigated how the hyperparameter *p* influences the model's translation efficacy, with *p* being the boundary that regulates the potency of visual similarity weights. Here are our findings: primarily, the data unequivocally show that our model perfects translation outcomes across many datasets with the hyperparameter *p* set up at 0.02. Furthermore, the data reveals a pronounced decline in BLEU and METEOR scores corresponding to incremental adjustments beyond or below the threshold *p*. We attribute this occurrence to dual principal factors. Firstly, a reduction in threshold p results in visual data capturing increased noise, which hurts model efficacy. Secondly, an elevation in threshold *p* results in the omission of potentially beneficial visual data, so diminishing the model's overall effectiveness.

Table 3: Ablation study on hyper-parameter ("*p*") on the EN→DE and EN→FR tasks

p	Multi30K EN→DE						Multi30K EN→FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	B	M	B	M	B	M	B	M	B	M	B	M
$p = 0$	40.94	57.16	34.02	53.91	29.82	48.23	61.07	76.02	54.11	72.09	44.13	65.23
$p = 0.01$	41.22	57.97	34.53	53.87	30.42	49.14	62.16	76.03	55.02	71.93	44.71	66.19
$p = 0.015$	41.31	58.51	34.96	54.33	30.75	49.63	62.73	76.81	55.17	72.09	45.32	66.17
$p = 0.02$	42.46	58.93	35.01	54.42	31.07	49.67	62.88	77.01	55.39	72.44	46.31	65.96
$p = 0.025$	41.07	57.39	34.03	54.01	31.04	49.26	62.17	76.04	54.83	71.96	46.03	67.32
$p = 0.03$	40.33	58.15	33.46	53.29	30.19	48.81	62.02	75.74	54.67	71.98	45.44	66.19

Ablation Study of Different Components of the Model: To assess the performance of each constituent part within our KDNR-MNMT model, we performed supplementary experiments and contrasted outcomes against various models, as detailed in Table 4. Our findings are as follows: (1) Effectiveness of the Bidirectional Knowledge Distillation Module: The experimental results show that removing the bidirectional knowledge distillation module results in a decline in both BLEU and METEOR metrics. This confirms that the bidirectional knowledge distillation module is an effective method for fusing multimodal features, aimed at enhancing translation performance. (2) Evaluating the Masked Cross-Modal Visual Encoder: To assess the module's ability to enhance performance, we used a gating mechanism for the integration of comprehensive visual and textual attributes. The experimental results prove that the model's performance decreases in the absence of the masking occlusion module. This result affirms our theory that the preemptive exclusion of extraneous visual data, followed by the integration of diverse modalities, positively changes the quality of translation results. (3) Effectiveness of the Novelly Designed Gating Fusion System: Compared to traditional gating fusion modules, the novel gating method we designed eases more effective fusion of cross-modal features.

Table 4: Presents the results of an ablation study examining various model components across the English-to-German and English-to-French translation tasks. KD_loss, NR_loss, and Gate_loss correspond to the experimental results of removing the knowledge distillation module, the masking occlusion module, and the gated fusion module from our model, respectively

Module	Multi30K EN→DE						Multi30K EN→FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	B	M	B	M	B	M	B	M	B	M	B	M
KDNR-MNMT	42.36	59.87	35.01	54.42	31.07	49.67	62.88	77.01	55.19	72.44	46.31	65.96
KD_Loss	42.27	59.03	34.76	54.11	30.46	48.23	62.18	76.37	54.76	71.68	45.97	65.04
NR_Loss	41.76	58.37	34.19	53.64	29.87	47.26	61.35	75.19	53.98	70.26	45.06	64.19
Gate_Loss	40.93	57.51	32.79	52.61	30.16	47.15	61.49	75.28	52.96	69.37	45.14	64.37

4.5 Case Study

As depicted in Table 5, our KDNR-MNMT model reaffirms the potency of an approach grounded in knowledge distillation and anti-noise interaction. The model iteratively cuts visual noise through knowledge distillation and masking occlusion, ensuring the best exploitation of feature information for machine translation. Traditional MNMT and NMT models, while capable of accurately making the principal content "schwarzer" from the image, struggle with the nuances such as "einen" and "im wasser," which, despite their minor visual presence, are pivotal for translation precision. We saw that visual noise inherent in the image data disrupts the MNMT model's functionality. Consequently, we engineered an encoder adept at getting more efficient representations, culminating in enhanced translation accuracy. This shows that by employing strategic two-way knowledge distillation and masking anti-noise interaction, our model can more effectively manage visual information, thereby bolstering the precision of machine translation.

Table 5: An example of the specific differences between our KDNR-MNMT and the traditional model in terms of EN→DE translation effects is given

	**src**	a black dog is retrieving a ball in water.
	tgt	ein schwarzer hund holt **einen** ball **im wasser**.
	MNMT	ein schwarzer hund holt einen bal **unterwasser**.
	NMT	ein schwarzer hund **apportiert** einen ball im wasser.
	Ours	ein schwarzer hund holt **einen** ball **im wasser**.
	**src**	the red car is ahead of the two cars in the background.
	tgt	das rote auto fährt vor **den beiden** autos im hintergrund.
	MNMT	das rote auto fährt vor **allen anderen** autos im hintergrund.
	NMT	das rote auto fährt vor **zwei** autos im hintergrund.
	Ours	das rote auto fährt vor **den beiden** autos im hintergrund.

4.6 Results on the EN→Cs Translation Task

We provide an in-depth validation of the effectiveness of the proposed method on the English to Czech (EN→CS) translation task, and the validation results are detailed in Table 6. Compared to all the control benchmarks, our model still exhibits excellent performance, which further confirms the effectiveness and generalizability of our model in dealing with different language pairs.

Table 6: Experiment results on Multi30K EN→CS task on BLEU and METEOR metrics

Model	Multi30K EN→CS			
	Test2016		Test2018	
	BLEU	METEOR	BLEU	METEOR
Transformer (NMT) [41]	32.70	32.34	27.62	29.03
Doubly-ATT [42]	33.25	32.28	29.12	29.87
Multimodal self-att [43]	33.12	32.01	28.75	29.51
Gated Fusion MNMT [44]	33.77	32.24	29.43	29.41
Our model	**34.91**	**33.46**	**30.93**	**30.07**

5 Conclusion

In this research, we introduce a pioneering approach to noise-resistant multimodal interaction fusion, integrating bidirectional knowledge distillation with a cross-modal relation-aware masking mechanism, tailored to mitigate noise in image features during multimodal fusion. Through comprehensive experimentation and thorough analysis of three authorized translation tasks, we substantiate the efficacy of our proposed KDNR-MNMT model, highlighting its marked performance superiority. Supplementary ablation studies reinforce the significance of bidirectional knowledge distillation in extracting analogous features and the strategic masking of irrelevant visual information, both of which positively influence the precision of machine translation. Moving forward, our agenda includes the pursuit of more refined techniques for noise elimination from visual data, with the goal of further enhancing the efficacy of multimodal machine translation systems.

Acknowledgement: This paper is supported by the Henan Provincial Science and Technology Research Project, the Science and Technology Innovation Project of Zhengzhou University of Light Industry.

Funding Statement: This study was supported by the Henan Provincial Science and Technology Research Project: 232102211017, 232102211006, 232102210044, 242102211020 and 242102211007, the Zhengzhou University of Light Industry Science and Technology Innovation Team Program Project: 23XNKJTD0205.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Erlin Tian, Fangmei Liu, Zuhe Li; data collection: Zengchao Zhu; analysis and interpretation of results: Zuhe Li, Zengchao Zhu; draft manuscript preparation: Zengchao Zhu. All authors reviewed the results and approved the last version of the manuscript.

Availability of Data and Materials: The data presented in this study are openly available in (Multi30k) at (https://arxiv.org/abs/1605.00459) (accessed on 21 January 2025) and (Flickr30k) at (http://shannon.cs.illinois.edu/DenotationGraph) (accessed on 21 January 2025), reference number [33,34].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Nam WY, Jang BC. A survey on multimodal bidirectional machine learning translation of image and natural language processing. Expert Syst Appl. 2024;235(1):121168. doi:10.1016/j.eswa.2023.121168.
2. Hou Z, Guo J. Virtual visual-guided domain-shadow fusion via modal exchanging for domain-specific multi-modal neural machine translation. In: MM '24: Proceedings of the 32nd ACM International Conference on Multimedia; 2024; AUS; p. 4227–35.
3. Li L, Tayir T, Han Y, Tao X, Velásquez JD. Multimodality information fusion for automated machine translation. Inf Fusion. 2023;91(1):352–63. doi:10.1016/j.inffus.2022.10.018.
4. Yin Y, Meng F, Su J, Zhou C, Yang Z, Zhou J, et al. A novel graph-based multi-modal fusion encoder for neural machine translation [M.S. dissertation]. China: Xiamen University; 2020.
5. Ye J, Guo J, Xiang Y, Tan K, Yu Z. Noise-robust cross-modal interactive learning with text2image mask for multi-modal neural machine translation. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022; Gyeongju, Republic of Korea. p. 5098–108.
6. Nishihara T, Tamura A, Ninomiya T, Omote Y, Nakayama H. Supervised visual attention for multimodal neural machine translation. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020; Sanya, China. p. 4304–14.
7. Li J, Ataman D, Sennrich R. Vision matters when it should: sanity checking multimodal machine translation models [M.S. dissertation]. Switzerland: ETH Zürich; 2021.
8. Tayir T, Li L, Tao X, Maimaiti M, Li M, Liu J. Visual pivoting unsupervised multimodal machine translation in low-resource distant language Pairs. In: Findings of the Association for Computational Linguistics: EMNLP 2024; 2024; Florida, USA; p. 5596–607.
9. Qin W, Sun T, Xiong D, Cui J, Wang B. Modeling homophone noise for robust neural machine translation. In: ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021; Toronto, UofT: CAD; p. 7533–7.
10. Yao S, Wan X. Multimodal transformer for multimodal machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; Washington, DC, USA; p. 4346–50.
11. Wang F, Yan J, Meng F, Zhou J. Selective knowledge distillation for neural machine translation [M.S. dissertation]. China: Peking University; 2021.
12. Hisamoto S, Post M, Duh K. Membership inference attacks on sequence-to-sequence models: is my data in your machine translation system? Trans Assoc Computat Linguist. 2020;8(1):49–63. doi:10.1162/tacl_a_00299.
13. Gain B, Bandyopadhyay D, Ekbal A. Experiences of adapting multimodal machine translation techniques for hindi [M.S. dissertation]. India: Indian Institute of Technology Patna; 2021.
14. Huang PY, Liu F, Shiang SR, Oh J, Dyer C. Attention-based multimodal neural machine translation. First Conf Mach Transl. 2016;2(1):639–45. doi:10.18653/v1/W16-23.
15. Calixto I, Liu Q, Campbell N. Incorporating global visual features into attention-based neural machine translation [M.S. dissertation]. Irish: ADAPT Centre Dublin City University; 2017.
16. Elliott D, Frank S, Barrault L, Bougares F, Specia L. Findings of the second shared task on multimodal machine translation and multilingual image description [M.S. dissertation]. England: School of Informatics University of Edinburgh; 2017.
17. Zhao Y, Komachi M, Kajiwara T, Chu C. Word-region alignment-guided multimodal neural machine translation. IEEE/ACM Trans Audio, Speech, Lang Process. 2021;30(1):244–59. doi:10.1109/TASLP.2021.3138719.
18. Song Y, Chen S, Jin Q, Luo W, Xie J, Huang F. Enhancing neural machine translation with dual-side multimodal awareness. IEEE Trans Multimed. 2021;24(1):3013–24. doi:10.1109/TMM.2021.3092187.
19. Lin H, Meng FD, Su JS, Yin YJ, Yang ZY, Ge YB, et al. Dynamic context-guided capsule network for multimodal machine translation. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020; Beijing, China. p. 1320–9.

20. Caruana R, Niculescu-Mizil A. Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2006; Philadelphia, USA. p. 535–41.
21. Hinton G. Distilling the knowledge in a neural network [Ph.D. dissertation]. USA: Google Inc.; 2015.
22. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: hints for thin deep nets [M.S. dissertation]. Spanish: Universitat de Barcelona; 2014.
23. Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2024; Honolulu, HI, USA. p. 3084–4000.
24. Gupta S, Hoffman J, Malik J. Cross modal distillation for supervision transfer. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, LV, USA. p. 2827–36.
25. Yuan M, Peng Y. Text-to-image synthesis via symmetrical distillation networks. In: MM '18: Proceedings of the 26th ACM international conference on Multimedia; 2018; London, LON: UK. p. 1407–15.
26. Song K, Tan X, Qin T, Lu JF, Liu TY. Mass: masked sequence to sequence pre-training for language generation. In: Proceeding of Machine Learning Research; 2019; California, CA, USA. p. 1–11.
27. Li P, Li L, Zhang M, Wu M, Liu Q. Universal conditional masked language pre-training for neural machine translation [Ph.D. dissertation]. China: Huawei Noah's Ark Lab; 2022.
28. Huang H, Liang Y, Duan N, Gong M, Shou L, Jiang D, et al. Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks [Ph.D. dissertation]. China: Microsoft Research Asia; 2019.
29. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
30. Bao HB, Li D, Wei FR, Wang WH, Yang N, Liu XD, et al. UniLMv2: pseudo-masked language models for unified language model pre-training. In: Proceedings of the 37th International Conference on Machine Learning; 2020; Hubei, China. p. 642–52.
31. Vaishali S, Neetu S. Enhanced copy-move forgery detection using deep convolutional neural network (DCNN) employing the ResNet-101 transfer learning model. Multimed Tools Appl. 2024;83(4):10839–63. doi:10.1007/s11042-023-15724-z.
32. Li ZW, Chen ZY, Yang F, Li W, Zhu YS, Zhao CY, et al. MST: masked self-supervised transformer for visual representation. Adv Neural Inf Process Syst. 2021;34(1):13165–76.
33. Elliott D, Frank S, Sima'an K, Specia L. Multi30k: multilingual english-german image descriptions [Ph.D. dissertation]. Holland: University of Amsterdam; 2016.
34. Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Computat Linguist. 2014;2(1):67–78. doi:10.1162/tacl_a_00166.
35. Mathur N, Baldwin T, Cohn T. Tangled up in BLEU: reevaluating the evaluation of automatic machine translation evaluation metrics [M.S. dissertation]. Australia: The University of Melbourne; 2020.
36. Hameed DA, Faisal TA, Alshaykha AM, Hasan GT, Ali HA. Automatic evaluating of Russian-Arabic machine translation quality using METEOR method. AIP Conf Proc. 2022;2386(1):040036. doi:10.1063/5.0067018.
37. Calixto I, Rios M, Aziz W. Latent variable model for multi-modal translation [M.S. dissertation]. Holland: The University of Amsterdam; 2018. LIUM-CVC Submissions for WMT17 Multimodal Translation Task [M.S. dissertation]. French: University of Le Mans; 2017.
38. Zhou M, Cheng R, Lee YJ, Yu Z. A visual attention grounding neural model for multimodal machine translation [M.S. dissertation]. Canada: University of California; 2018.
39. Caglayan O, Madhyastha P, Specia L, Barrault L. Probing the need for visual context in multimodal machine translation [M.S. dissertation]. French: Le Mans University; 2019.
40. Wang D, Xiong D. Efficient object-level visual context modeling for multimodal machine translation: masking irrelevant objects helps grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021; New York, NY, USA. p. 2720–8.
41. Vaswani A. Attention is all you need [Ph.D. dissertation]. USA: Google Research; 2017.

42. Wu Z, Kong L, Bi W, Li X, Kao B. Good for misconceived reasons: an empirical revisiting on the need for visual context in multimodal machine translation [M.S. dissertation]. China: The University of Hong Kong; 2021.
43. Gong H, Jia M, Jing L. Multimodal interaction modeling via self-supervised multi-task learning for review helpfulness prediction [M.S. dissertation]. China: Sun Yat-sen University; 2024.
44. Ye J, Guo J, Yu Z. The progressive alignment-aware multimodal fusion with easy2hard strategy for multimodal neural machine translation. In: Proceedings of ICLR; 2023; Kijali, KGL, Rwandan.
45. Ji B, Zhang T, Zou Y, Hu B, Shen S. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective [M.S. dissertation]. China: Tencent Minority-Mandarin Translation; 2022.
46. Ye J, Guo J. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. Appl Intell. 2022;52(12):14194–203. doi:10.1007/s10489-022-03331-8.