Multimodal Neural Machine Translation: A Survey of the State of the Art

Yi Feng[1], **Chuanyi Li**[1*], **Jiatong He**[1], **Zhenyu Hou**[1], **Vincent Ng**[2]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Human Language Technology Research Institute, University of Texas at Dallas, USA
{fy,lcy}@nju.edu.cn, vince@hlt.utdallas.edu

Abstract

Multimodal neural machine translation (MNMT) has received increasing attention due to its widespread applications in various fields such as cross-border e-commerce and cross-border social media platforms. The task aims to integrate other modalities, such as the visual modality, with textual data to enhance translation performance. We survey the major milestones in MNMT research, providing a comprehensive overview of relevant datasets and recent methodologies, and discussing key challenges and promising research directions.

1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015; Gehring et al., 2017; Zhu et al., 2024) is the task of translating the source language into the target language using neural networks. In practical scenarios, however, text is often accompanied by data from other modalities. For instance, in scenarios such as cross-border e-commerce, social media, and news reporting, the text frequently co-occurs with multi-view images. The accompanying visual information typically carries valuable language-agnostic information, which can be leveraged to enhance and complement the corresponding semantic interpretation. In fact, Caglayan et al. (2016) have indicated that integrating additional modalities with the textual modality can effectively improve translation quality, especially for ambiguous, gender-related and domain-specific complex content. This finding broadens the scope of text-based NMT into a multimodal paradigm, providing a novel view to enhance the performance of machine translation, referred to as Multimodal Neural Machine Translation (MNMT).

The MNMT task, which was formally introduced in the WMT 2016 Shared Task (Specia et al., 2016), aims to translate the source sentence into the target sentence by incorporating additional modalities, most commonly in the form of visual information. MNMT research has primarily focused on two scenarios: (1) image-text semantic alignment, where the accompanying image semantically aligns with the text, and (2) image-text semantic noise[1], where the image is irrelevant to the text. Figure 1 illustrates the results of a NMT model and a MNMT model on an example sampled from a cross-border e-commerce platform under these two scenarios.

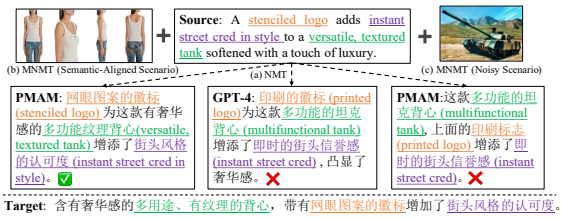

Figure 1: Example of MNMT, with outputs produced by PMAM for semantically aligned (left) and noisy scenarios (right). For comparison purposes, the NMT output produced by GPT-4 (center) is shown.

In this example, the expressions "stenciled logo", "instant street cred in style" and "versatile, textured tank", according to professional e-commerce translators, should be translated as "网眼图案的徽标", "街头风格的认可度", and "多用途、有纹理的背心". Among the three model outputs, only the output of PMAM (Guo et al., 2024a), a MNMT model, is correct in the semantically aligned scenario. In particular, even GPT-4 (Achiam et al., 2023), a text-based large language model, incorrectly translates these expressions as "印刷的徽标 (printed logo)", "即时的街头信誉感 (instant street cred)" and "多功能的坦克背心 (multifunctional tank)". Such mistranslations could undermine consumer engagement and enthusiasm on cross-border e-commerce platforms, potentially harming overall user experi-

[*]Corresponding author

[1]Noisy images can be commonly found in daily life (e.g., on social media/e-commerce platforms), and are often caused by low-quality visuals or misleading interpretations due to ambiguity.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22130–22147

ence and trust. Therefore, research into the MNMT task is of paramount importance.

Unlike existing reviews of MNMT (Sulubacak et al., 2020; Gwinnup and Duh, 2023; Nam and Jang, 2024; Shen et al., 2024), this paper centers on broadening the application scope of MNMT. Amid paradigm shifts (from specialized compact models to cross-scale collaboration, and from text-image co-modality to generalized multimodal systems), it synthesizes core challenges and proposes novel research directions, thereby expanding the research landscape of MNMT and diversifying future avenues. More specifically, we make three key contributions in this survey. First, from a practical perspective, this work introduces novel modeling challenges, which we identify as particularly relevant to the future directions of MNMT. Second, reflecting the aforementioned paradigm shift in MNMT from general to domain-specific tasks (e.g., e-commerce, law), this paper proposes a three-way categorization of existing MNMT datasets (general-domain, domain-specific, and multi-domain) and identifies their limitations. Finally, the survey distills future research directions for MNMT, proposing numerous novel and promising avenues.

2 Modeling Challenges and Issues

In this section, we discuss seven modeling challenges and issues for MNMT researchers.

1. Visual imbalance. In real-world scenarios, the textual modality is often accompanied by several distinct situations: (1) one-to-many text-image correspondence (i.e., a text corresponds to multiple images); (2) one-to-one text-image correspondence (i.e., a text corresponds to a single image); and (3) text-only scenarios (i.e., a text appears without any accompanying images). Consequently, several core challenges arise: (1) How can we effectively handle multiple images and leverage their interrelationships to enhance translation quality when a single text corresponds to several images? (2) In cases where only one image is available, how can we extract semantic-related visual representations that correspond to the source sentence to improve translation performance? (3) In the absence of images, how can we adaptively hallucinate semantic-related visual features linked to the source sentence while ensuring effective fusion of the textual and visual modalities? As shown in Figure 1, the example includes four images that present different angles of the source sentence. However, existing MNMT models face challenges in adaptively handling an arbitrary number of images without preprocessing. For instance, PMAM, one of the models used to produce the outputs in Figure 1, exploits only the image that is most semantically related to the text despite the fact that four images are present.

2. Semantic shift across different scenarios. Visual information can serve as a double-edged sword. On one hand, the integration of semantically-aligned images could enhance translation performance. On the other hand, the integration of noisy images may result in generated sentences that deviate from the original semantics, a problem known as *semantic shift*. As depicted in Figure 1, the MNMT model that employs semantically-aligned images produces accurate translations of the target sentence. Conversely, when the model incorporates noisy-image information, it generates erroneous translations such as "多功能的坦克背心 (multi-functional tank top)", "印刷标志 (printed logo)", and "即时的街头信誉感 (instant street cred)", which deviate significantly from the original meaning. Therefore, a challenge lies in designing a MNMT model that can adapt to various scenarios. Specifically, in semantically aligned visual contexts, it should fully leverage the complementarity between text and images to enhance translation quality, but in noisy environments, it must filter out irrelevant information to maintain the accuracy of translations.

3. Computation costs and latency of MNMT tasks. Successful deployment of MNMT models depends on not only their translation performance, but also the deployment costs and computational latency. In existing MNMT models, effective translation typically comes at the expense of a substantial increase in the number of model parameters, significantly raising deployment costs and latency. For instance, the length of the visual features extracted from popular architectures such as Resnet-101 (He et al., 2016), Faster R-CNN (Ren et al., 2016) and Vision Transformers (Dosovitskiy et al., 2021) are notably longer than those extracted from the corresponding textual modality. Moreover, if multiple stacked vision encoder blocks are employed, the number of parameters can far exceed that of the textual modality. How to reduce the number of visual information parameters without degrading translation performance remains a key challenge.

4. Sense disambiguation. Ambiguous textual expressions often encompass a range of semantic interpretations, potentially eliciting entirely opposite

representations in different contexts. Examples include multi-sense words, gender-related phrases, and domain-specific fine-grained expressions. As illustrated in Figure 1, the multi-sense word "tank", the domain-specific fine-grained expression "stenciled logo", and the fashionable jargon "instant street cred in style" exemplify ambiguous expressions that pose significant challenges for translation in the absence of the corresponding images. In fact, existing MNMT models primarily concentrate on disambiguation in common scenarios, which typically exhibit simple expressions and singular contexts, such as common descriptions of daily life or casual conversations, exemplified by phrases like "Two dogs bark behind a fence" (an example from the Multi30k dataset). In these instances, model design often overlooks the corresponding visual information, failing to leverage the supplementary role of images in the translation process effectively. However, in certain complex scenarios, such as the example in Figure 1, relying solely on textual information often fails to convey the intended meaning accurately. In such cases, visual information plays a crucial role in translating specific expressions. How to effectively leverage visual data for sense disambiguation remains a significant challenge.

5. Unsupervised multimodal neural machine translation. The integration of the visual modality with the textual modality not only enhances translation performance for MNMT, but also provides a novel perspective for addressing challenges in unsupervised multimodal neural machine translation (UMNMT). The images could serve as language-agnostic information, playing a crucial role in bridging linguistic gaps when a parallel corpus is absent. Furthermore, they can facilitate connections between languages, turning the unsupervised learning problem into a supervised one. However, existing UMNMT research primarily focuses on widely used languages, such as English, Chinese, and French, leaving the challenge of low-resource UMNMT largely unaddressed. Consequently, a significant challenge concerns how to fully leverage the pivotal role of visual information to enhance machine translation performance in low-resource languages.

6. Model interpretability. There is little work on building interpretable MNMT models that can *explain* how a model arrives at the translated sentence. Naturally, the explanation would be in the form of a natural-language paragraph that explains step by step how the different modalities (e.g., text and images) are combined to eventually arrive at the translated sentence. Building interpretable MNMT models can help us better understand the role of the visual modality in MNMT. There are currently two perspectives on the role of the visual modality in MNMT. One camp hypothesizes that visual information is limited and often redundant for MNMT, and its introduction functions as a form of regularization, resulting in marginal improvements to performance (Wu et al., 2021; Elliott, 2018). The other camp posits that visual information can effectively complement textual information when the text is incomplete or insufficient (Delbrouck and Dupont, 2017; Caglayan et al., 2019; Li et al., 2021, 2022a; Wu et al., 2021). Having explanations could provide empirical support for either of these camps.

7. Adaptations of LLMs and MLLMs for MNMT. The impressive language understanding and generation capabilities of Large Language Models (LLMs), along with the outstanding performance of Multimodal Large Language Models (MLLMs) in handling multimodal data, have attracted increasing attention from researchers. However, existing LLMs cannot process multimodal information, whereas MLLMs cannot be directly applied to MNMT tasks without adaptations as they require precise, context-specific alignment between textual and visual data, which often exceeds the alignment capabilities of standard MLLMs. For instance, as depicted in Figure 1, given the source sentence, GPT-4 inaccurately generates phrases such as "多功能的坦克背心 (multifunctional tank)" and "即时的街头信誉感 (instant street cred)", and fails to capture significant details, such as "网眼图案的徽标 (stenciled logo)". These results highlight a deficiency in the text generation capabilities for LLMs, particularly when they are given complex expressions that are complemented with visual information. Therefore, a critical challenge lies in how to combine LLMs' linguistic capabilities with MLLMs' multimodal processing capabilities across visual, auditory, and other modalities to advance MNMT tasks.

3 Corpora

3.1 Dataset Categorization

Table 1 compares the commonly-used MNMT corpora along six dimensions: (1) ambiguity, (2) balance, (3) language, (4) domain, (5) size, and (6) additional annotations (if any). These corpora belong to one of the following three categories.

Dataset	Sub-Dataset	Ambiguous	Balance	Language	Domain	Text	Image	Video	Additional Annotation
Flicker	Flicker8K Flicker30K	✗	✗	EN	-	40,460 158,915	8,092 31,783	- -	Lack of non-English parallel corpus
MS COCO	-	✗	✗	EN	-	1.6M	328K	-	-
WIT	-	✗	✓	108 languages	multiple domains	37.6M	11.5M	-	Non-English and low-resource languages
Conceptual Captions	CC 3M CC 12M	✗	✓	EN	multiple domains	3.3M 12.4M	3.3M 12.4M	- -	-
IAPR TC-12	-	✗	✓	EN, DE	general domains	20,000	20,000	-	-
Multi30K	-	✗	✓	EN, DE, FR, CS	general domains	31,014	31,014	-	-
MLT	-	✓	✓	EN, DE EN, FR	general domains	53,868 44,779	53,868 44,779	- -	4-tuples contain ambiguous words
MultiSense	-	✓	✓	EN, DE, ES	general domains	9,504	9,504	-	Verb sense ambiguity
AmbigCaps	-	✓	✓	EN, TR	general domains	91,601	91,601	-	Gender ambiguity
M^3	M^3-Multi30K M^3-AmbigCaps	✗ ✓	✓ ✓	CS, DE, EN, - FR, HI, LV, TR	general domains	31,014 91,601	31,014 91,601	- -	-
TIT Dataset	-	✗	✓	EN, ZH ZH, EN EN, DE	-	1.0M 1.0M 1.0M	1.0M 1.0M 1.0M	- - -	-
BLATID	-	✗	✓	EN, ZH	-	1.2M	1.2M	-	-
OCRMT30K	-	✗	✓	EN, ZH	-	164.7K	30.2K	-	-
Fashion-MMT	Fashion-MMT(L) Fashion-MMT(C)	✗	✗	EN, ZH	specific domains	114,257 40,000	885,244 312,656	- -	The text / image ratio is 0.129 The text / image ratio is 0.128
EMMT	-	✗	✗	EN, ZH	specific domains	875.0K	22.0K	-	-
How2	-	✗	✓	EN, PT	general domains	191.6K	-	191.6K	-
VaTeX	-	✗	✗	EN, ZH	general domains	412.7K	-	41.3K	-
BigVideo	-	✗	✓	EN, ZH	general domains	4.5M		4.5M	-
VISA	VISA-Polysemy VISA-Omission	✓	✓	EN, JA	general domains	20.7K 19.2K	- -	20.7K 19.2K	-

Table 1: Comparison of several widely used corpora for MNMT. The "Ambiguous" column indicates whether the dataset has been deliberately rendered ambiguous, such as through intentional selection of polysemous nouns or verbs or by incorporation of a specific threshold of polysemous words. The "Balance" column pertains to whether the dataset demonstrates modality-specific imbalance.

General-domain MNMT datasets are often built by pairing images with multilingual text descriptions to enhance translation quality, especially in resolving ambiguous expressions. For example, Multi30k (Elliott et al., 2016), an extension of the Flickr30k (Young et al., 2014) dataset, includes descriptions in multiple languages. The translations were performed by professional translators without showing the original images, simulating real-world translation scenarios.

To address the challenges related to ambiguity, several datasets have been proposed. MLT (Lala and Specia, 2018) focuses on testing modality alignment, MultiSense (Gella et al., 2019) explores cross-modal mapping of ambiguous words, and AmbigCaps (Li et al., 2021) investigates how the visual modality aids in resolving gender ambiguity.

Some datasets encode the relationship between video content and the corresponding text, capturing dynamic scenes and details to improve the multimodal processing capabilities of MNMT models. For example, How2 (Sanabria et al., 2018) ensures that each step in its instructional videos is reflected in text descriptions, including timing and procedural details; VaTeX (Wang et al., 2019) captures dynamic video information and generates multilingual subtitles; and VISA (Li et al., 2022d) uses visual context to address translation ambiguities.

Domain-specific datasets are more complex, as accurately translating domain-specific concepts is challenging, especially when dealing with domain-related terms, slang, or jargon. Therefore, the construction of these datasets often requires professional experts to verify the accuracy of the content and optimize the semantic alignment across modalities. For instance, Fashion-MMT (Song et al., 2021) focuses on multimodal fusion in e-commerce scenarios to improve the translation of domain-specific expressions. Meanwhile, EMMT (Zhu et al., 2023) focuses on aligning the visual and textual modalities and resolving ambiguity, and investigates how visual information influences the semantic interpretation of complex text, particularly when dealing with polysemous words and ambiguous contexts, to improve translation quality.

Multi-domain datasets cover a broad range of topics and applications, supporting models' generalization across different contexts. These datasets originate from various industries and platforms, such as social media, e-commerce, and encyclopedias, with diverse languages, cultures, and visual content. For example, WIT (Srinivasan et al., 2021) extracts large-scale multilingual image-text pairs from Wikipedia, covering topics like history, sci-

ence, and culture. The Conceptual Captions dataset (Sharma et al., 2018; Changpinyo et al., 2021) automatically extracts and refines textual descriptions from Web content and advertisements, ensuring that the descriptions align with the semantic features of the images.

3.2 Dataset Limitations

Below we discuss the limitations of these datasets.

Long tail phenomenon and lack of transferability in MNMT models. Most MNMT datasets exhibit the following limitations. First, they lack diverse scenarios. Their textual content usually comprises brief and straightforward sentences that often describe everyday scenarios, such as "a small black dog jumping over gates". The paired images are similarly simple in terms of semantics. The absence of complex syntactic structures and rich linguistic expressions in both the textual and visual modalities limits MNMT models' ability to handle intricate translation tasks. Second, they lack professional terminology or domain-specific representations. Many of them predominantly focus on general expressions, with very limited coverage of professional terminology or domain-specific representations. For instance, terms such as "myocardial infarction" in the medical field or "contract breach" in the legal sector are seldom represented in these datasets. Finally, they overlook differences in cultural representations. Specifically, they often overlook the metaphoric and cultural nuances of language, which are crucial across various contexts and domains. For example, phrases like "time is money" in English or "一箭双雕 (killing two birds with one stone)" in Chinese involve a deep understanding of culture and context. These complexities could limit a model's transferability and accuracy across new domains or different scenarios.

Scarcity of multimodal data in low-resource contexts. In MNMT, the scarcity of multimodal data in low-resource contexts poses a significant challenge. Current datasets predominantly focus on mainstream languages such as Chinese, English, and French, often ignoring languages with large speaker populations but limited resources, such as Bengali and Vietnamese. This scarcity not only constrains the broader application of MNMT models but also undermines their generalization capabilities across diverse linguistic contexts.

Cross-modal adversarial samples and robustness deficiencies. MNMT datasets mainly focus on standard text-image alignment scenarios but often ignore the importance of adversarial contexts. For instance, replacing a semantically-aligned image with a noisy image confuses the model, as shown in Figure 1. These issues could be attributed to dataset construction bias. Specifically, during construction, most MNMT datasets tend to select text-image samples with semantic correspondence. While this approach facilitates the model's ability to learn cross-modal correlations, the lack of adversarial samples could make a model vulnerable in noisy scenarios.

Lack of textual-visual semantic co-occurrences. In MNMT datasets, the relationship between text and images often only reflects superficial and direct associations, failing to delve into more complex semantic co-occurrences. Although an image may display multiple objects, the corresponding textual description might only mention a subset of them. For instance, in the CoMMuTE dataset (Futeral et al., 2023), a typical example might describe "We'll have to get rid of that mole", while the associated image could also show other background elements like pedestrians passing by or surrounding trees. This partial correspondence results in a significant amount of visual information that is not mentioned in the text becoming noise, thereby complicating the model's ability to effectively capture the deeper semantic relationships across the two modalities and failing to use these visual cues.

We conclude this section by noting that these limitations are identified through a rigorous analysis of the inherent shortcomings in existing corpora, all of which represent critical challenges requiring prioritized attention in MNMT research.

4 Approaches to MNMT

In this section, we divide existing approaches to MNMT into five categories, which roughly correspond to the modeling challenges outlined in Section 2.[2] Unlike previous classification schemes, our categories adopt an application-oriented perspective, establishing novel categorical dimensions rooted in real-world applications.

4.1 Disambiguation-based Approaches

To address Challenge 4, disambiguation-based approaches are designed to exploit multiple modalities for disambiguating ambiguous expressions.

[2]The state-of-the-art models for each approach are discussed in Table 4 in Appendix A.

Approach		Method	Description
Disambiguation via multimodal fusion		Cross-modal gating	introduces gating mechanisms to selectively regulate the contribution of the visual and textual modalities, enabling the model to prioritize the modality with the most relevant information for disambiguating context-sensitive terms effectively (Ye et al., 2022; Ye and Guo, 2022; Guo et al., 2023a; Li et al., 2022a; Yin et al., 2020, 2023; Cheng et al., 2024; Hou and Guo, 2024).
		Cross-modal attention	actively establishes semantic interactions between the visual and textual modalities by attending to the most relevant features across modalities, thereby enhancing the model's ability to disambiguate and generate more accurate translations (Ive et al., 2019; Lin et al., 2020; Guo et al., 2024a; Su et al., 2019; Tayir et al., 2024; Yao and Wan, 2020; Wang and Xiong, 2021; Yang et al., 2020; Zhao et al., 2021; Caglayan et al., 2019).
Disambiguation via model training	Cross-modal training	Masking	masks important text information, forcing models to use visual information for inference (Caglayan et al., 2021; Song et al., 2021; Yang et al., 2024b; Futeral et al., 2023, 2025).
		Text-image matching	learns the matching relationship between text and image, thereby enabling the model to better understand the semantic connections between textual-visual modalities and enhancing the alignment of cross-modal information (Song et al., 2021; Yang et al., 2024b).
		Classification	makes predictions on region-level image patches, which allows a model to learn fine-grained cross-modal feature representations and thereby enhances its understanding of image details and achieves more precise semantic alignment during the translation process (Caglayan et al., 2021).
	Multi-task learning	Visual-guided tasks	help the model better understand and extract visual information, such as grounded representation prediction (Elliott and Kádár, 2017), visual agreement regularized training (Yang et al., 2020), and image captioning tasks (Futeral et al., 2023; Cheng et al., 2023).
		Multimodal-oriented tasks	focus on improving the model's ability to associate information across different modalities, such as cross-modal feature alignment (Hou and Guo, 2024; Zhou et al., 2018), object-masking (Wang and Xiong, 2021), and triplet alignment (Peng et al., 2022b) between the source language, the target language, and the paired images.
Detaching Images in the Testing Phase		Visual hallucination	hullucinates in order to eliminate the dependency on the image modality and enhance model robustness (Li et al., 2022c; Peng et al., 2022a; Calixto et al., 2019; Yuasa et al., 2023).
		Adaptive image selection	retrieves visual representations related to the source language semantics, thereby eliminating the dependency on the image modality and enhancing the model's adaptability and robustness (Long et al., 2024; Wang et al., 2024; Tang et al., 2022; Fang and Feng, 2022; Zhang et al., 2020).
Visual-Balance		Single-view visual fusion	enhances translation by capturing semantic features from a single image, making them suitable for one-to-one image-text correspondence (Guo et al., 2023a; Li et al., 2022a; Yin et al., 2020; Lin et al., 2020; Wang and Xiong, 2021; Guo et al., 2024b; Li et al., 2022b).
		Text-guided generation	relies on diffusion models (Calixto et al., 2019; Yuasa et al., 2023; Guo et al., 2023b) to generate visual representations related to the source language semantics, thereby improving translation quality for text-only scenarios (Calixto et al., 2019; Yuasa et al., 2023; Guo et al., 2023b).
Visual-Pivot		Supervised MNMT	trains models on aligned multimodal data in which the visual modality not only provides contextual information for source-target translation but also functions as a language-pivot (Yuasa et al., 2023; Guo et al., 2024b; Yang et al., 2024a).
		Unsupervised MNMT	uses images as a bridge (Tayir and Li, 2024; Fei et al., 2023; Huang et al., 2020; Li et al., 2022b), employing techniques like scene graph generation (Fei et al., 2023) or multimodal prompts (Yang et al., 2024a) to capture latent semantic connections between languages and the visual modality, even without aligned labels.

Table 2: Description of the methods employed by four of the five approaches to MNMT.

4.1.1 Disambiguation via Information Fusion

Designing effective cross-modal fusion mechanisms to integrate textual and visual information for resolving ambiguity has become a research focus in MNMT. Current cross-modal fusion mechanisms can be divided into (1) cross-modal gating and (2) cross-modal attention, as discussed in Table 2. Despite having promising results, these methods tend to rely overly heavily on visual information and are not robust to real-world scenarios where the images can be noisy.

4.1.2 Disambiguation via Model Training

Cross-modal pre-training. Cross-modal pre-training tasks are designed to enhance models' joint understanding of the textual and visual modalities, particularly through leveraging visual representations to resolve textual ambiguities. Several cross-modal pre-training methods have emerged, including (1) masking, (2) text-image matching, and (3) classification, as discussed in Table 2.

Multi-task learning. MNMT studies have employed joint training methods to integrate visual

and textual information. These multi-task learning approaches can be divided into two categories: (1) visual-guided tasks and (2) multimodal-oriented tasks, as discussed in Table 2. Despite significant success in cross-modal alignment and disambiguation through aligned image-text pairs, existing methods tend to over-rely on such strict alignments. This reliance limits a model's ability to generalize effectively, especially when faced with diversity and the long-tail phenomena in real-world scenarios or specialized domains.

4.2 Visual-Balance Approaches

Recall that several issues surround Challenge 1: (1) How can we effectively handle multiple images and leverage their interrelationships to enhance translation quality when a single text corresponds to several images? (2) In cases where only one image is available, how can we extract semantic-related visual representations that correspond to the source sentence to improve translation performance? (3) In the absence of images, how can we adaptively hallucinate semantic-related visual features linked to the source sentence while ensuring effective fusion of textual and visual modalities?

Visual-balance approaches to MNMT have predominantly focused on the latter two scenarios, and can be categorized into two types: (1) single-view visual-guided fusion approaches and (2) text-guided image-generation approaches, as discussed in Table 2. In contrast, research on many-to-one image-text correspondence is scarce.

4.3 Visual-Pivot Approaches

With its language-agnostic nature, visual information can highlight the commonalities among different languages. Hence, images serve as both a bridge and a pivot modality in translation tasks, especially in unsupervised and low-resource language translation scenarios. To address Challenge 5, MNMT researchers have developed vision-based pivot translation methods, including both supervised and unsupervised MNMT methods (see Table 2). While unsupervised methods have shown potential in low-resource unsupervised MNMT tasks, current visual-pivot methods still fail to fully exploit the cross-lingual bridging role of the visual modality in machine translation. Furthermore, their computational complexity exceeds that of traditional MNMT approaches due to the requirement for complex feature processing operations (e.g., disentanglement, pooling) on the visual modality.

4.4 LLM/MLLM-Based Approaches

To address Challenge 7, researchers have combined LLMs and MLLMs for MNMT, transforming visual representations into formats that can be understood by language models to enhance translation performance (Vijayan et al., 2024; Gupta et al., 2023; Futeral et al., 2025). For example, Futeral et al. (2025) employ NLLB (Costa-jussà et al., 2022) models as the base translation model, which is then adapted into a MNMT model by adding lightweight trainable modules, enabling the use of visual information for disambiguation.

How well do LLMs and MLLMs perform on MNMT? According to recent studies (Futeral et al., 2025; Gupta et al., 2023; Zuo et al., 2023; Zhu et al., 2024), some observations can be made. For *general-domain* datasets, such as Multi30k, the performance of GPT-4 is on par with that of current state-of-the-art models (Zhu et al., 2024). In contrast, other LLMs, such as Qwen-7B-chat, perform poorly, lagging significantly behind GPT-4. When MLLMs (e.g., Qwen-VL-chat) are used, performance deteriorates significantly. We hypothesize that the performance drop could be attributed to the inability of MLLMs to effectively align fine-grained image-text representations. The visual modality, as noise, disrupts the translation process. In *domain-specific* translation scenarios, LLMs generally perform poorly and often result in semantic shifts, as shown in Figure 1. Similarly, when MLLMs (e.g., GPT-4o) are used for translating domain-specific text with semantically aligned images and text, the translation results are often worse than those of LLMs. We speculate that this deficiency may be attributed to the phenomenon of "visual hallucination".

4.5 Approaches for Detaching Images in Testing

Some attempts have focused on how to decouple the visual modality during the testing phase to overcome the data format limitations of MNMT tasks, which require triplet data (source sentence, target sentence, and the corresponding images). Current image-independent approaches can be categorized into two types: (1) visual hallucination and (2) adaptive image selection, as discussed in Table 2. While these methods focus on overcoming the limitations of the triplet data, they overlook another closely-related challenge, semantic shift, which occurs when the model is exposed to noisy visual con-

Method	Category	Description
Evaluating text translation quality	N-gram-based	compares the generated and reference translations based on word n-grams (e.g., BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005)) and character n-grams (e.g., chrF (Popović, 2015)).
	edit distance-based	calculates the minimum number of editing operations needed to modify machine translations to match reference translations, such as TER (Snover et al., 2006) and WER (Morris et al., 2004).
	pre-trained model-based	derives evaluation metrics based on pre-trained models, which better capture semantic quality, such as COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020).
Evaluating contribution of visual information	cross-modal gating mechanism	adjusts the integration of visual and textual data, revealing the model's dependency on visual input and providing insights into cross-modal interactions (Cheng et al., 2024; Tayir et al., 2024).
	cross-modal attention mechanism	computes the attention weights between the textual and visual modalities to capture visual representations that align with the textual semantics (Ive et al., 2019; Yao and Wan, 2020; Caglayan et al., 2019).
	visual-textual semantic alignment	assesses the similarity between textual and visual representations using techniques such as cosine similarity to verify the accuracy of visual information and the effectiveness of integration (Yang et al., 2020; Zhao et al., 2021; Liu et al., 2021; Fei et al., 2023).
	evaluating visual contextual reasoning	assesses the model's ability to reason with and effectively utilize visual information in specific contexts for translation tasks (Zuo et al., 2023; Li et al., 2022a).
	evaluating visual information absence and impact	quantifies the contribution of visual information to the translation task by observing the model's performance difference when visual information is missing (Wang and Xiong, 2021; Long et al., 2024; Delbrouck et al., 2017).
	adversarial evaluation	evaluates the effectiveness of visual information by testing the model's performance with congruent and incongruent visual information (Elliott, 2018; Cheng et al., 2023).
Evaluating cross-modal fusion	info gain evaluation	quantifies the effect of modality fusion on translation performance (Ji et al., 2022).
	modal difference calculation	reflects the level of information sharing between modalities by quantifying the differences between them, thereby assessing their collective impact on translation quality (Futeral et al., 2025; Hou and Guo, 2024).

Table 3: Details of the three automatic evaluation methods for MNMT.

texts, leading to a shift in meaning. As depicted in Figure 1, when the model incorporates noisy-image information (i.e., image of the tank), it generates erroneous translations such as "多功能的坦克背心 (multi-functional tank top)", "印刷标志 (printed logo)", and "即时的街头信誉感 (instant street cred)".

5 Evaluation Issues

5.1 Evaluation Methods

Automatic evaluation. Existing automatic evaluation methods for MNMT can be broadly divided into three categories: (1) methods for evaluating the text translation quality of MNMT models; (2) methods for evaluating the contribution of visual information to the performance of MNMT models; and (3) methods for evaluating the effectiveness of fusing the information from the visual and textual modalities. Details of these methods, as well as the corresponding metrics, are discussed in Table 3.

Human evaluation. Human evaluation has been used to complement automatic evaluation metrics for evaluating MNMT outputs that involve fluency and adequacy assessment (Zhao et al., 2021) and quality estimation (Specia et al., 2010).

5.2 Evaluation Challenges

While many automatic metrics have been used, there is currently a lack of a *standard set* of evaluation metrics. This makes it difficult to track progress in the field, since the reliance on different metrics in different papers could make it impossible to directly compare MNMT models w.r.t. translation quality and the role of visual information. The situation is further aggravated by the fact that there is also a lack of a standard set of evaluation datasets: as discussed before, numerous datasets have been developed in the past eight years. The key challenge, therefore, involves designing a standardized evaluation framework that would facilitate the automatic comparison of different models.

6 Ethical Considerations

In this section, we discuss the ethical considerations associated with MNMT research.

Data privacy and security. MNMT datasets typically include user-generated text and images, which may contain identifiable personal details

such as faces, license plates, and potentially even geographic locations. Thus, proper data cleaning and anonymization are critical to prevent personal information leaks. Additionally, encryption technologies should be employed to protect data during transmission and storage to prevent unauthorized access and data leaks.

Fairness and bias. Dataset biases include those that involve gender, race, and culture. These biases may be learned by the model and reflected in the translation results. This can not only affect the quality of translations but also lead to discriminatory practices when these models are applied in real-world scenarios. To mitigate biases, datasets should extensively cover various geographical regions, languages, and cultural backgrounds.

Transparency. Since MNMT architectures are often black-box systems, the internal mechanisms remain obscure. Enhancing model transparency means making the model's operating mechanisms, decision-making basis, and limitations known to users. For example, developing visualization tools that illustrate how the model utilizes image information and how it integrates multimodal information to enhance machine translation quality could enhance model transparency.

Cultural sensitivity and adaptability. In real-world scenarios, the content requiring translation may involve sensitive topics specific to certain cultures or social groups. Consider the source sentence "A woman is wearing a headscarf while praying in the mosque". In this example, the "headscarf" holds specific religious significance in Muslim culture. Misinterpreting it as a "hat" may appear disrespectful toward religious and cultural values, and in certain contexts, it could even cause offense. Therefore, when handling content related to religion, culture, or social groups, MNMT models must be sensitive to these expressions and accurately preserve their cultural meaning.

7 Concluding Remarks

We conclude with directions for MNMT research.

Corpora. Existing MNMT datasets have several limitations, including limited scene diversity, the lack of domain-specific terms, and insufficient representation of metaphors and cultural context. These issues hinder the full exploration of the visual modality, affecting its role in different scenarios. Moreover, the lack of explanations for how to combine the modalities to produce the correct translation makes it non-trivial to develop interpretable MNMT models in a supervised manner. We recommend that efforts be devoted to the development of annotated corpora for MNMT that address as many of the aforementioned limitations as possible. Details can be found in Appendix B.1.

Evaluation. To facilitate the comparison of different MNMT models, we recommend the development of a standard evaluation framework that should be composed of a set of evaluation datasets and metrics that all MNMT researchers should use. We believe that the development of such a framework would benefit the field in the long run, but it would require the consensus of the community. We therefore recommend that MNMT researchers discuss this issue and develop a shared vision for the field through a shared task on MNMT. Our suggestion to develop a standard evaluation framework is by no means an attempt to discourage development of new metrics/datasets: as better metrics are developed, we can incorporate them into the framework. Details can be found in Appendix B.6.

Modeling. Given our categorization of existing approaches to MNMT, we identified several key weaknesses of MNMT models. Specifically, MNMT models (1) are not interpretable; (2) fail to produce fine-grained alignment of the text-image modalities needed for accurate translation; (3) struggle with semantic shifts across different scenarios; (4) are not effective at capturing cross-modal interactions, particularly those that involve abstract entities and concepts such as sentiment, sarcasm, and other pragmatic aspects; (5) improve translation quality at the expense of an explosion in the number of model parameters; (6) focus largely on MNMT for rich-resource languages; and (7) fail to effectively leverage state-of-the-art LLMs and MLLMs.

We recommend the development of MNMT models that can address one or more of these weaknesses. Details on how to design effective cross-modal fusion mechanisms and those that can handle abstract entities and concepts can be found in Appendix B.2 and Appendix B.4 respectively. For suggestions on how to incorporate LLMs and MLLMs into MNMT tasks, we refer the reader to Appendix B.7. As far as how to handle semantic shifts across different scenarios, we propose to enhance MNMT models with collaborative multi-agent systems (see Appendix B.5 for details).

Limitations

Since virtually all MNMT systems incorporate only the vision modality, our discussion has primarily focused on vision-text modalities, without discussing other modalities like audio or sensor data.

Acknowledgments

We thank the reviewers for their valuable comments on an earlier draft of this paper. This work was supported by National Natural Science Foundation of China (No. 62406139), State Key Laboratory for Novel Software Technology at Nanjing University (KFKT2025A15, ZZKT2025B14, KFKT2024A07, ZZKT2024B02).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of th 3rd International Conference on Learning Representations*, San Diego, CA, USA.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, Virtual.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Soul-mix: Enhancing multimodal machine translation with manifold mixup. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11283–11294, Bangkok, Thailand. Association for Computational Linguistics.

Xuxin Cheng, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023. DAS-CL: Towards multimodal machine translation via dual-level asymmetric contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 337–347, Birmingham, United Kingdom.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Stéphane Dupont, and Omar Seddati. 2017. Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation. *arXiv preprint arXiv:1707.01009*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Sylvain Gelly Georg Heigold, Jakob Uszkoreit, and Neil Houlsby:. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, Virtual.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, Toronto, Canada. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2025. Towards zero-shot multimodal machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 761–778, Albuquerque, New Mexico. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjun Guo, Zhenyu Hou, Yantuan Xian, and Zhengtao Yu. 2024a. Progressive modality-complement aggregative multitransformer for domain multi-modal neural machine translation. *Pattern Recognition*, 149:110294.

Junjun Guo, Rui Su, and Junjie Ye. 2024b. Multi-grained visual pivot-guided multi-modal neural machine translation with text-aware cross-modal contrastive disentangling. *Neural Networks*, 178:106403.

Junjun Guo, Junjie Ye, Yan Xiang, and Zhengtao Yu. 2023a. Layer-level progressive transformer with modality difference awareness for multi-modal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3015–3026.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023b. Bridging the gap between synthetic and authentic images for multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2863–2874, Singapore. Association for Computational Linguistics.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. CLIPTrans: Transferring visual knowledge with pre-trained models for multimodal machine translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2875–2886, Paris, France.

Jeremy Gwinnup and Kevin Duh. 2023. A survey of vision-language pre-training from the lens of multimodal machine translation. *arXiv preprint arXiv:2306.07198*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA.

Zhenyu Hou and Junjun Guo. 2024. Virtual visual-guided domain-shadow fusion via modal exchanging for domain-specific multi-modal neural machine translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne, VIC, Australia. Association for Computing Machinery.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu, and Si Shen. 2022. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6755–6764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. 2022b. Video pivoting unsupervised multi-modal machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3918–3932.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022c. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5216–5226, New Orleans, LA, USA.

Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022d. VISA: An ambiguous subtitles dataset for visual scene-aware machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6735–6743, Marseille, France. European Language Resources Association.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329, Virtual. Association for Computing Machinery.

Xiao Liu, Jing Zhao, Shiliang Sun, Huawen Liu, and Hao Yang. 2021. Variational multimodal machine translation with underlying semantic alignment. *Information Fusion*, 69:73–80.

Zi Long, ZhenHao Tang, Xianghua Fu, Jian Chen, Shilong Hou, and Jinze Lyu. 2024. Exploring the necessity of visual modality in multimodal machine translation using authentic datasets. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 36–50, Torino, Italia. ELRA and ICCL.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings of the 8th International Conference on Spoken Language Processing*, pages 2765–2768, Jeju Island, Republic of Korea.

Wongyung Nam and Beakcheol Jang. 2024. A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Systems with Applications*, 235:121168.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ru Peng, Yawen Zeng, and Jake Zhao. 2022a. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2379–2390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ru Peng, Yawen Zeng, and Junbo Zhao. 2022b. Hybrid-Vocab: Towards multi-modal machine translation via multi-aspect alignment. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 380–388, Newark, NJ, USA.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges. *arXiv preprint arXiv:2405.12669*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Product-oriented machine translation with cross-modal cross-lingual pretraining. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2843–2852, Virtual. Association for Computing Machinery.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24:39–50.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, Virtual. Association for Computing Machinery.

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, Long Beach, CA, USA.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147.

ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022. Multimodal neural machine translation with search engine based image retrieval. In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Turghun Tayir and Lin Li. 2024. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–22.

Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024. Encoder-decoder calibration for multimodal machine translation. *IEEE Transactions on Artificial Intelligence*, 5(8):3965–3973.

Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Adding multimodal capabilities to a text-only translation model. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–28, Chicago, USA. Association for Machine Translation in the Americas.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2720–2728, Virtual. AAAI Press.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, Seoul, Republic of Korea.

Yan Wang, Yawen Zeng, Junjie Liang, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2024. RetrievalMMT: Retrieval-constrained multi-modal prompt learning for multi-modal machine translation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 860–868, Phuket, Thailand. Association for Computing Machinery.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6153–6166, Online. Association for Computational Linguistics.

Jian Yang, Hongcheng Guo, Yuwei Yin, Jiaqi Bai, Bing Wang, Jiaheng Liu, Xinnian Liang, Linzheng Cahi, Liqun Yang, and Zhoujun Li. 2024a. m3P: Towards multimodal multilingual translation with multimodal prompt. *arXiv preprint arXiv:2403.17556*.

Jian Yang, Hongcheng Guo, Yuwei Yin, Jiaqi Bai, Bing Wang, Jiaheng Liu, Xinnian Liang, LinZheng Chai, Liqun Yang, and Zhoujun Li. 2024b. m3P: Towards multimodal multilingual translation with multimodal prompt. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10858–10871, Torino, Italia. ELRA and ICCL.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9418–9425, New York, NY, USA. AAAI Press.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Junjie Ye and Junjun Guo. 2022. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*, 52(12):14194–14203.

Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022. 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*, pages 5098–5108, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Yongjing Yin, Jiali Zeng, Jinsong Su, Chulun Zhou, Fandong Meng, Jie Zhou, Degen Huang, and Jiebo Luo. 2023. Multi-modal graph contrastive encoding for neural machine translation. *Artificial Intelligence*, 323:103986.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal neural machine translation using synthetic images transformed by latent diffusion model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, and Mingxuan Wang. 2023. Beyond triplet: Leveraging the most data for multimodal machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2679–2697, Toronto, Canada. Association for Computational Linguistics.

Yuxin Zuo, Bei Li, Chuanhao Lv, Tong Zheng, Tong Xiao, and JingBo Zhu. 2023. Incorporating probing signals into multimodal machine translation via visual question-answering pairs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14689–14701, Singapore. Association for Computational Linguistics.

A State-of-the-Art MNMT Systems

Table 4 discusses the strengths and weaknesses of the state-of-the-art MNMT systems on five commonly used evaluation datasets.

	System	Approach	Description	Strengths	Weaknesses
WIT&MULTI30K	Gupta et al. (2023)	Image-independent, LLM-based	The paper integrates the semantic depth of LLMs with the visual perception capabilities of MLLMs. By designing a lightweight mapping network, it converts visual representations into text tokens and concatenates them with the target language text, thereby providing the decoder with rich contextual information.	(1) This model combines the advantages of LLMs and MLLMs. (2) It incorporates a lightweight mapping network that converts visual representations into a format suitable for processing by language models. (3) This model does not require additional image dependencies and does not need image inputs during the inference phase, making it more flexible and efficient for practical applications.	(1) Training and fine-tuning large pre-trained models like CLIP and mBART require significant computational resources. (2) The model's performance is contingent upon the pre-training data, which limits its effectiveness in low-resource or underrepresented languages. (3) The model's fairness and accuracy across diverse cultural contexts may be affected, necessitating particular attention to data bias issues.
M^3	Guo et al. (2022)	Disambiguation-based	The model proposed in this paper utilizes a pre-trained visual encoder and a Transformer-based encoder to integrate visual and textual information. Additionally, it introduces a language-aware prompt generation module (LVPG) to dynamically generate visual prompts for different target languages.	(1) The model achieves dynamic adaptation of visual information across different languages by generating visual prompts related to the target language, thereby enhancing cross-lingual translation effectiveness. (2) LVP-M3 employs a co-attention strategy to effectively integrate textual and visual information during the encoding and translation processes, enabling the model to better capture the semantic interactions between text and images. (3) By utilizing dynamically generated visual prompts, LVP-M3 enhances translation performance in specific domains, particularly in situations where visual information can help eliminate ambiguities or capture contextual details.	(1) The model is highly dependent on high-quality image-text alignment data, and its performance may decline when the data is incomplete or contains noise. (2) Generating high-quality visual prompts requires significant computational resources, limiting the model's scalability in multilingual and diverse data scenarios. (3) In low-resource languages or with limited data, the model may struggle to effectively generate language-aware visual prompts, thereby impacting translation performance. (4) The visual prompts are closely tied to specific languages, and introducing new languages or domains may lack flexibility, leading to suboptimal performance of the model in unforeseen scenarios.
FASHION-MMT	Guo et al. (2024a)	Disambiguation-based	Due to the significant gap between visual and textual modalities, direct interaction can lead to modality collapse. This paper proposes a bidirectional progressive cross-modal interaction mechanism that effectively narrows the gap between images and text. It introduces a cross-modal interaction-based modality-complementary multi-Transformer (BPMCT) to extract domain-relevant multimodal representations, thereby significantly improving the translation quality of domain representations.	(1) The cross-modal interaction-complementarity mechanism effectively narrows the modality gap between text and images through a bidirectional progressive modality complementarity approach. (2) The model employs multi-layer modality interactions to more comprehensively capture complementary information from multimodal data. (3) Through the Cross-Modal Adaptive Fusion (CAF) module, the model can adaptively integrate visual and textual information based on a multimodal gating mechanism, thereby enhancing translation accuracy and robustness.	(1) There exists a gap between the visual and textual modalities, and direct interaction can lead to modality collapse, meaning the effective integration of information from both modalities becomes challenging. (2) The alignment requirements between the visual and textual modalities are stringent, making it difficult to obtain suitable data in practical applications. (3) The model relies on semantically aligned data, and its performance may decline in noisy environments.
EMMT	Hou and Guo (2024)	Disambiguation-based, Image-independent	This paper introduces a domain shadow fusion method guided by virtual visual scenes, utilizing adaptive distillation and a modality swapping mechanism to achieve a simpler multimodal interaction framework. Additionally, it explores the significance of transitional modalities in the cross-modal distillation process.	(1) The modality mixing selection voting strategy aids in integrating dispersed visual details from the domain, aggregating modality-mixed domain representations and text, thereby enhancing the collaboration between domain features and textual semantics. (2) Guided by virtual visual scenes, the model generates smoother multimodal representations, helping to reduce the representation gap between modality-mixed domain shadow details and the original text. (3) The cross-modal swapping mechanism allows for the aggregation of dispersed visual details across modalities, facilitating the integration of domain-specific multimodal information.	(1) Although the model performs well in specific domains, maintaining consistent performance across multiple domains remains a challenge, necessitating reliance on meticulous alignment strategies. (2) In scenarios lacking domain-relevant visual data, the model's performance may be limited, reducing its applicability. (3) Despite model compression, the multi-step fusion process still requires substantial computational resources, limiting its scalability. (4) Effectively exchanging domain-specific features between different modalities is complex, and improper handling may lead to the loss of critical semantic information.

Table 4: Strengths and weaknesses of the state-of-the-art MNMT systems on five commonly used evaluation datasets.

B Discussion of Future Directions

B.1 Developing New Datasets for MNMT

Current MNMT datasets have several limitations, including limited scene diversity, lack of domain-specific terms, and insufficient representation of metaphors and cultural context. For example, in datasets like Multi30k and WMT, where textual information is abundant, the visual modality often serves a regularization role. In contrast, in domain-specific or multi-sense datasets like Fashion-MMT, EMMT, and 3AM, visual information plays a crucial role in semantic supplementation, helping models capture important visual features for more accurate translations. To advance MNMT research and better integrate text and visual modalities, developing more comprehensive and diverse datasets is essential.

In addition to general requirements such as the construction of large-scale datasets, coverage of multiple and low-resource languages, diversification of scenes and coverage of specialized expressions, MNMT tasks should also fulfill the following characteristics:

1. Increasing the complexity of image-text correspondence: Enabling each text to correspond to any number of visual images enhances the model's ability to understand information from various perspectives and improves its adaptability to diverse real-world scenarios.

2. Design of adversarial samples: Testing with adversarial samples can effectively enhance the robustness and accuracy of the model.

3. Integration of rich cultural and contextual elements: By incorporating multimodal data with specific cultural backgrounds and contexts, the sensitivity and robustness of the model to cultural differences and contextual variations can be enhanced.

B.2 Designing Effective Cross-Modal Fusion Mechanisms

Existing MNMT works can be categorized into two types based on the need for visual information. The first type, where images are essential for translation, occurs typically in scenarios with limited or incomplete textual contexts. This is particularly relevant when addressing ambiguities or complex expressions, as visual information provides crucial semantic support. The second type, where images are optional, usually occurs when the textual information is sufficient, and visual data acts more like regularization, offering limited improvement to translation performance.

One of the core reasons for the aforementioned classification is that current multimodal fusion frameworks struggle to effectively align and integrate fine-grained visual-text features. Specifically, the use of visual information exhibits a dual-edged effect: on one hand, visually aligned information serves as an important semantic supplement or regularization, improving translation quality; on the other hand, noisy visual information may disrupt the translation process, leading to incorrect translation as shown in Figure 1. Consequently, for a cross-modal fusion mechanism to be effective, it should be designed to dynamically identify and capture text-related visual information and remain robust against noisy images. For example, the mechanism should ideally adjust the weight of the visual information automatically by real-time analysis of the semantic consistency between the visual and text data, thereby ensuring translation accuracy and adaptability to different contexts.

B.3 Interpretability of Visual Effectiveness

Since the introduction of MNMT tasks, the effectiveness of visual information has been a topic of ongoing debate. As mentioned in Section B.2, MNMT tasks can be categorized into two primary scenarios based on the varying requirements for visual information. For general domain translation tasks, numerous studies suggest that visual information mainly serves a regularization role. In contrast, in domain-specific tasks or when important semantic information is missing from the text, visual information typically acts as a supplementary modality.

However, these studies have several limitations, including: (1) inconsistency in evaluation frameworks; (2) incomplete dataset designs, where visual information often appears redundant; and (3) underdeveloped cross-modal semantic alignment and fusion mechanisms. As a result, existing MNMT methods often offer a narrow evaluation of visual effectiveness, failing to comprehensively explore its underlying mechanisms. Therefore, we recommend that researchers focus more on the scientific exploration of the role of visual information.

B.4 Capturing Cross-Modal Sentiment-Aware Aspects for MNMT

MNMT tasks have demonstrated broad potential across various application scenarios, particularly

Text: Oh great, it's raining again. Just what I needed on a day I planned to sunbathe. #BestDayEver

Figure 2: An example of sarcasm sampled from an international social media platform.

within international social platforms. Consider the sample in 2, which showcases a sarcastic post drawn from an international social media platform. In this example, the combination of textual and visual information constructs a classical sarcastic expression, which both the textual content and the imagery collaboratively convey a deeper meaning that is contrary to the literal interpretation. Furthermore, the utilization of the hashtag "#BestDayEver" further enhances the extent of sarcastic tone, demonstrating that the day is far from the "great" description provided. Therefore, to enhance the practicality of MNMT tasks in such application scenarios, future research should focus on analyzing emotional expressions and figurative language in both images and text to achieve accurate translations of such expressions.

B.5 Enhancing MNMT with Collaborative Multi-Agent Systems

In global news reporting scenarios, existing MNMT methods often struggle with semantic shifts across different scenarios when processing rapidly changing multi-source data, such as real-time videos, social media updates, and official text reports. These issues significantly affect the deployment performance of MNMT models. With the rapid development of multi-agent systems, they are capable of efficiently collaborating in a multi-task environment and can effectively process and integrate information from various modalities. Each agent focuses on processing data from a specific source, collaborating to produce unified and accurate translation results.

However, despite the potential advantages, the application of multi-agent systems to MNMT tasks faces several challenges and limitations:

1. Quality issues in continuous translation: Continuous translation tasks often result in outputs that do not meet expectations, including unnecessary explanations and loss of critical information such as scenario details.

2. Cross-linguistic and cultural adaptability: MNMT must handle multimodal content from diverse languages and cultures. Agents in a multi-agent system typically specialize in processing specific types of data, lacking the flexibility to adapt to semantic and emotional differences across cultures, especially in scenarios with insufficient training data.

3. High computational cost: Multi-agent systems, when executing complex MNMT tasks, generally require significant computational resources, which substantially increases computational costs.

B.6 Constructing a Unified Evaluation Framework

As described in Section B.2, the classification of MNMT tasks based on visual information highlights an unresolved issue: existing MNMT methods generally lack a comprehensive framework for evaluating both text translation quality and the role of visual information, particularly in terms of accurately assessing the impact of visual information on translation quality. This limitation prevents a deeper understanding of the visual modality's function in translation tasks, highlighting the need for an evaluation framework that can assess the influence of visual information both quantitatively and qualitatively. For instance, the methodology of metamorphic testing can be adopted by applying specific visual modifications to observe corresponding changes in translation results. During the training phase, specific noise can be added to image features to analyze its impact on the contextual understanding in translations. Additionally, when the background of images changes, it is crucial to evaluate whether the translation maintains semantic consistency and accuracy.

However, there still remains some limitations when applying metamorphic testing methods in MNMT tasks. The difficulties include:

1. Complex semantic alignment: In MNMT tasks, the semantic alignment between modalities such as text and images is complex and variable, complicating metamorphic testing's ability to encompass all potential semantic relationships and nuances.

2. Cultural and contextual dependency: Metamorphic testing may not accurately assess the appropriateness or accuracy of translations without sufficient contextual information.

3. Dynamic content challenges: Metamorphic testing must rapidly adjust its strategies to handle new data from real-time updates or dynamic multimodal content. Traditional metamorphic testing often lacks the necessary flexibility and adaptability to manage rapid changes in dynamic contexts effectively.

B.7 Exploiting MLLMs and LLMs for MNMT

In recent years, with the development of LLMs and MLLMs, an increasing number of researchers have focused on utilizing these technologies to address challenges in MNMT tasks. According to recent studies (Futeral et al., 2025; Gupta et al., 2023; Zuo et al., 2023; Zhu et al., 2024), two observations can be made:

First, for general-domain datasets, such as Multi30k, the performance of GPT-4 is on par with current state-of-the-art models (Zhu et al., 2024). In contrast, other LLMs, such as Qwen-7B-chat, perform poorly, lagging significantly behind GPT-4. When MLLMs, such as Qwen-VL-chat, are used, model performance deteriorates significantly, likely due to the introduction of the visual modality. We hypothesize that the core issue lies in the current inability of MLLMs to effectively align fine-grained image-text representations. The visual modality, as noise, disrupts the translation process.

Second, in domain-specific translation scenarios, language models generally perform poorly and often result in semantic shifts, as shown in Figure 1. Similarly, when multimodal large models (e.g., GPT-4o) are used for translating domain-specific text with semantically aligned images and text, the translation results are often worse than those of language models. We speculate that this deficiency may be attributed to the phenomenon of "visual hallucination". Therefore, it is evident that while LLMs perform well in general-domain translation tasks, they struggle with domain-specific expressions. Additionally, MLLMs may encounter issues like visual hallucinations, indicating a deficiency in understanding visual information, leading to translations that lack domain-specific features.

Based on these findings, we propose the following future research directions:

1. Cost-effective optimization of large models: Although current large models excel on general-domain datasets, this performance is often achieved by increasing the model parameters, leading to challenges in training costs and deployment. We suggest that MNMT researchers focus on "downsizing" large models through methods like distillation and transfer learning, transferring the strong translation capabilities of large models to smaller models. This not only integrates the advantages of both large and small models but also reduces training costs and significantly improves translation quality, thus enhancing deployment capabilities in real-world applications.

2. Collaborative use of large and small models: In real-world scenarios, translation tasks often vary in complexity. For different translation tasks, we should effectively assign tasks based on their difficulty, using large models for complex tasks and small models for simpler ones. This not only saves computational resources but also enhances practical application capabilities.

3. Optimizing fine-grained alignment and fusion of image-text modalities in MLLMs: Current research on multimodal large models faces several challenges, including visual hallucination, visual noise, and the fine-grained semantic alignment of image and text modalities. The introduction of visual information may confuse multimodal large models, leading to inaccurate translations and semantic drift. Therefore, optimizing the alignment and fusion of image-text modalities remains a critical challenge in future MNMT research.