

Article

Visual Semantics in MT Evaluation: Do Image Descriptions Help with Assessment of Multimodal MT Quality?

Sami Ul Haq ^{1,*} , Sheila Castilho ¹  and Yvette Graham ² 

¹ ADAPT Centre, School of Applied Language and Intercultural Studies, Dublin City University, D09 V209 Dublin, Ireland; sheila.castilho@dcu.ie

² ADAPT Centre, Trinity College Dublin, D02 PN40 Dublin, Ireland; ygraham@tcd.ie

* Correspondence: sami.haq2@mail.dcu.ie

Abstract

Multimodal machine translation (MMT) aims to integrate visual context with textual data to improve the translation of ambiguous source text, such as the inclusion of an image as additional context. However, the evaluation of systems largely still relies on automatic metrics designed to evaluate text alone, and do not account for additional modalities during evaluation. The lack of dedicated MMT evaluation methods often results in inconsistent findings and creates uncertainty regarding the actual contribution of visual context in translation. In this work, we examine the performance of state-of-the-art trained and untrained evaluation metrics, particularly when comparing multimodal and text-only systems. Our evaluation focuses on the degree to which existing metrics are sensitive enough to distinguish between multimodal and text-only machine translation systems. We further investigate the potential for automatically generated image descriptions to serve as effective contextual signals for improving metric sensitivity to multimodal tasks. Our results show that incorporating such visual information into supervised metrics yields better alignment with human judgment. While all metrics successfully distinguished image-aware from image-agnostic systems on general test sets, both n-gram-based and embedding-based metrics struggled with respect to contrastive evaluation designed to capture context-dependent errors. Furthermore, we discuss how the presence of visual context may influence human evaluator judgment, observing that given the opportunity, human ratings are often substantially revised, further emphasizing the critical role of context in the evaluation of MMT.

Keywords: MMT evaluation; automatic evaluation; context-aware evaluation



Academic Editor: Arkaitz Zubiaga

Received: 17 May 2026

Revised: 5 July 2026

Accepted: 6 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Machine Translation (MT) has achieved remarkable progress in recent years, largely driven by advances in neural text-generation architectures and large-scale pre-training [1]. End-to-end encoder–decoder models enabled learning broader representations instead of relying on phrase-based statistical alignments, allowing Neural Machine Translation (NMT) systems to capture long-range dependencies and contextual semantics more effectively. Developments in data augmentation, pre-training, and transfer learning have further facilitated multilingual NMT approaches [2], extending to enable a single model to handle numerous language pairs. These methods allow high-resource languages to enhance performance in low-resource settings and support more generalised translation capabilities.

Despite these advances, MT systems continue to struggle with linguistic ambiguity [3] and producing contextually appropriate translations [4], particularly in cases where

text alone is insufficient to resolve meaning [5]. Words with multiple senses, pronouns referring to visually grounded entities, or culturally situated expressions often require extra-linguistic information for accurate interpretation. To address this challenge, Multimodal Machine Translation (MMT) has emerged as a research direction that integrates visual features—typically from paired images—alongside textual input to provide additional disambiguating context [6–8].

While architectural progress in MMT has been substantial, automatic evaluation methodologies to track these advances have lagged behind [4]. Commonly used automatic evaluation metrics (AEMs) such as BLEU [9], METEOR [10], and COMET [11] are largely image-agnostic, assessing translation quality primarily through lexical overlap or semantic similarity. As a result, they fail to capture the visual grounding capabilities that are central to MMT [12]. In addition, benchmark test sets often consist of short and syntactically simple sentences [13], which further limits the ability of existing metrics to reliably reflect system quality in multimodal settings. Consequently, the lack of modality-aware evaluation frameworks has led to ongoing uncertainty about whether reported improvements in MMT systems can be attributed to the inclusion of visual information [14]. The absence of standardised evaluation metrics and datasets further complicates cross-system comparison, making it difficult to measure progress in the field when different studies adopt different evaluation protocols [4].

In principle, image context should be incorporated directly into MMT evaluation, in a similar manner as human assessors rely on visual cues to resolve ambiguities in translation judgments [12]. Related work in image captioning evaluation demonstrates that metrics integrating visual features align better with human judgments [15–17], highlighting the importance of multimodal grounding for assessment. However, extending these approaches to MMT is non-trivial for two main reasons. First, in image captioning the image and caption are intrinsically aligned, whereas in MMT the image often serves only as an auxiliary cue for resolving ambiguity and may not correspond directly to all aspects of the textual content. Second, translation ambiguity is inherent to the MT task itself, meaning that evaluation methods must assess not only semantic similarity but also consistency and adequacy across languages.

In parallel, pre-trained AEMs have demonstrated improved performance through the use of document-level or multi-sentence context, enabling better modelling of discourse coherence and coreference resolution [18–20]. These metrics operate either at the sentence level by augmenting embedding representations with surrounding context or at the document level by computing similarity scores across extended text spans [21,22]. Given their ability to integrate extra-sentential information and discourse signals, exploring the applicability of such metrics for MMT evaluation constitutes a promising direction.

In this paper, we investigate and compare multimodal and text-only MT systems through the lens of state-of-the-art AEMs, with a particular focus on the reliability of text-only metrics in multimodal settings. We propose a simple approach to incorporating visual semantics into existing metrics by using target-language image descriptions (IDs) as a proxy for visual context. Through a comparative evaluation of context-aware and context-agnostic metrics on multilingual multimodal benchmark datasets, we aim to address the following research question (RQ):

- RQ: Can image descriptions, when integrated with existing pre-trained AEMs, improve contextual sensitivity and evaluation accuracy in MMT?

The remainder of this paper is organised as follows. Section 2 provides an overview of MT, multimodal MT, and their evaluation. Section 3 reviews existing MMT evaluation methods and discusses their limitations. Section 4 introduces our methodology for integrating IDs into metric-based evaluation. Section 5 describes the experimental setup, followed

by the meta-evaluation in Section 6. Finally, Section 7 concludes this paper by discussing how different AEMs capture different aspects of MMT system performance.

2. Background

2.1. Machine Translation

MT is the task of automatically translating text from one language into another. MT models aim to capture and model the complex linguistic features of languages—such as syntax, semantics, and pragmatics—that are fundamental to understanding language and generating accurate translations. Since its inception, MT research has undergone significant evolution, transitioning from rule-based approaches to statistical methods [23,24], advancing to state-of-the-art NMT systems [25–27], and now entering the emerging era of Large Language Models (LLMs) [28]. Modern MT systems, powered by vast amounts of bilingual data and advanced computational resources, can produce translations of considerable quality [29].

NMT currently dominates the MT field due to its simplified translation model, faster data processing, and more efficient memory utilization compared to earlier paradigms [30]. These models typically employ encoder–decoder neural network architectures: the encoder processes the source sentence as input, while the decoder generates the target sentence conditioned on the encoder’s output. NMT systems based on the Transformer architecture have achieved state-of-the-art performance [27].

While statistical and neural models have traditionally relied on labelled data for training, recent advances in pre-training have enabled the modelling of sequential problems through self-supervised and multi-task learning paradigms, as exemplified by LLMs [1]. These models, capable of understanding and generating human-like text, can perform a wide range of Natural Language Processing (NLP) tasks. The emergence of LLMs introduces new opportunities for understanding how knowledge is represented and transferred across languages, domains, and modalities, and how these models may redefine the scope and methodology of MT [31].

2.2. Multimodal MT

MMT extends traditional MT with the integration of information from multiple modalities—such as text, images, video, or audio—to enhance the accuracy and contextual understanding of translation output. MMT aims to augment the latent representations derived from standard translation systems with information drawn from these additional modalities [7]. The fundamental intuition driving MMT is that visual context can resolve translation ambiguities and supply missing textual information, thereby leading to improved translation quality over text-only MT. This research area gained prominence through a series of dedicated shared tasks held at the Conference on Machine Translation (WMT) (<https://statmt.org> (accessed on 20 June 2026)) starting in 2016, primarily focusing on image-guided translation [6,8,32].

The growing interest in MMT is largely driven by the increasing availability of diverse multimodal datasets and the architectural flexibility of neural models, which can jointly process multiple modalities such as vision, speech, and text [2,33]. These non-textual modalities can provide helpful context, particularly in cases where MT systems struggle—such as the translation of ambiguous words whose correct interpretation and rendering depend heavily on contextual information [13,34]. For instance, the English word “bank” might translate differently in French depending on whether the image shows a financial institution (“banque”) or a river edge (“rive”). Similarly, an image can clarify whether the English word “glasses” should be translated as ‘drinking vessels’ or ‘spectacles’, as depicted in the example given in Figure 1. MMT systems can leverage the visual input to make these

disambiguation decisions [35], as the visual context can aid in modelling complex grammatical features, such as gender marking (e.g., inferring the sex of a person in the image to correctly translate from a gender-neutral to a gendered language). Caglayan et al. [7] and Futral et al. [13] highlight the robustness of MMT models against input noise, such as spelling mistakes or deliberately missing input words, suggesting that visual context can mitigate deficiencies when linguistic context is scarce [7,13].

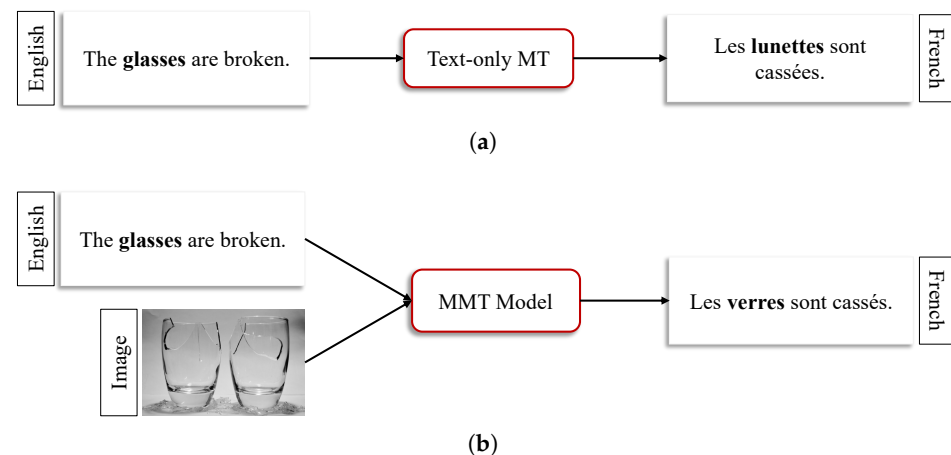


Figure 1. An example showing the difference between the conventional text-only MT and MMT models: (a) The incorrect translation of the polysemous word “glasses” into eyeglasses (“lunettes”) by the text-only MT model. (b) The MMT model correctly translated the polysemous word into drinking glasses (“verres”) with the help of a relevant image.

MMT systems are primarily image-guided models built on the frameworks of neural networks [7,36], integrating visual signals in various architectural stages [13,37,38]. These methods are generally classified into early-fusion and late-fusion strategies. Early fusion aims for a closer, often pre-encoder, integration of textual and visual representations, typically utilizing global visual features (e.g., averaged pooled ResNet-50 features) [8,12,32], such as through encoder–decoder initialization in Recurrent Neural Networks (RNNs) [39–42]. Other early techniques include treating the visual feature vector as a pseudo-word, and prepending or appending it to the source embeddings, or using element-wise multiplication [43]. Advanced early fusion approaches, like the multimodal Transformer [44], integrate multimodal self-attention layers to fuse modality representations deeply within the encoder. Conversely, late fusion strategies rely on separate encodings of modalities and a merge of representations later, usually closer to the decoder, often leveraging hierarchical attention [45] and gating mechanism [14,40]. This involves an auxiliary visual attention layer over the image data to produce a distinct visual context vector [5], which is subsequently combined with the textual context using summation, concatenation [46], or sometimes specialised mechanisms like hierarchical attention [45,47]. Furthermore, other fusion and training paradigms exist, such as Multi-task Learning (MTL), which jointly trains the text encoder to predict the visual feature vector [35,48], or modern approaches like ZeroMMT [38] and VGAMT [13], which adapt pre-trained text models via lightweight adapters and a visually conditioned, masked language modelling (VLM) objective combined with a Kullback–Leibler (KL) divergence penalty to maintain translation fidelity [38].

In another growing area of research, large-scale vision–language models (LVLMs) [15,49,50] have demonstrated strong zero-shot and few-shot performance across a range of downstream multimodal tasks (e.g., image caption, text-to-image, or image-to-text generation), often achieving impressive results with minimal or no task-specific data [16,51]. Models like CLIP [49], Flamingo [52], and BLIP [50] have recently been investigated for their potential

visually informed translation capabilities [53]. Notably, ref. [16] aligns features from a multilingual text model with those extracted by CLIP using a lightweight mapping network, while [54] combines visual pre-training with cross-lingual masked language modelling and region classification to produce visually grounded representations that, when fine-tuned, outperform unimodal MT baselines.

The above discussion highlights both the growth and potential of MMT systems; however, a key obstacle remains, namely the lack of standardised evaluation methods for reliably and accurately tracking research progress in this area. Current evaluation practices continue to rely heavily on string-based metrics such as BLEU, METEOR, and CHRF, often with little or no complementary analysis involving expert human assessment. As a result, there is ongoing debate about whether visual context contributes meaningful information beyond what is already available in the text [14].

2.3. MT Evaluation

Evaluation constitutes a fundamental component of MT research, serving as both a diagnostic tool and a benchmark for measuring system progress. Broadly, evaluation methods can be classified into human and automatic approaches.

2.3.1. Human Evaluation

Human evaluation (HE) remains the most reliable method for assessing translation quality, as it directly reflects human perceptions of fluency, adequacy, and contextual appropriateness. When conducted by professionals, HE has the advantage of capturing linguistic nuances, pragmatic meaning, and discourse coherence—dimensions that automated methods may fail to consider. Over the years, several HE metrics have been proposed, with adequacy and fluency remaining the most common criteria, typically assessed using various scales [55–57] and performed by either human experts or crowd-workers [58,59]. Metrics based on error typologies, such as MQM [60] and Error Span Annotation (ESA) [61], are considered more reliable than simple scoring methods. MQM requires human evaluators to explicitly identify and annotate errors in the translation output rather than assigning a single score to a segment or sentence [62], while ESA offers a lighter error identification metric, similar to Direct Assessment (DA) [58].

Post Editing (PE), in addition to its role in computer-aided translation and revision, is another widely used approach for MT evaluation [63] as PE effort is often interpreted as an indirect measure of MT quality. In semi-automatic settings, where MT output is corrected by a human translator, the Human Translation Edit Rate (HTER) score—calculated as the ratio of edits to the total number of words—serves as an estimate of quality, with lower scores indicating better quality translations [64]. HE also plays a fundamental role in advancing automatic evaluation research, as many evaluation metrics rely on human-annotated data to train neural models and/or to align their scoring behaviour with human judgments.

However, despite its interpretative richness, human evaluation is inherently time-consuming, costly, and susceptible to subjective variation [57,65]. Moreover, inter-annotator inconsistencies [66], differences in linguistic proficiency, and variations in task design can all affect the reliability and reproducibility of the evaluation. Although human evaluation remains essential, there are scenarios in which relying solely on human judgments is impractical, such as when rapid feedback is required during system development or when a large number of models must be compared. In such cases, MT evaluation is typically performed as a two-stage process, where automatic metrics are used for initial system filtering and parameter optimization due to their efficiency, and scalability, while human evaluation is reserved for the final assessment of system quality [67].

2.3.2. Automatic Evaluation

Automatic evaluation (AE) provides a scalable and cost-effective alternative to HE, allowing for rapid benchmarking during system development and optimization. Early heuristic-based metrics, most notably BLEU [9], TER [64], METEOR [10], and ChrF [68], offered a simple approach by quantifying n-gram overlaps between candidate and reference translations. Such methods provided an efficient proxy for translation quality and were instrumental in advancing statistical and early neural models. However, as MT systems have evolved—particularly with the advent of NMT—the limitations of these metrics have become increasingly evident [69]. Word- and n-gram-based measures tend to penalise legitimate paraphrasing, alternative phrasal reordering, or stylistic variation, and often perform poorly on morphologically rich languages [69–71]. In response, the field has seen a transition towards embedding-based and pre-trained metrics [72,73], which utilise distributed semantic representations from language models such as BERT [72]. These approaches provide a more nuanced understanding of meaning beyond surface-level matching, yet they remain susceptible to biases inherent in training data, domain sensitivity, and cross-lingual inconsistencies [74,75]. To overcome limitations of LM-based pre-trained metrics, a more recent trend has become fine-tuning models on translations with human evaluation scores [11,69,76,77]. Most of these metrics are either n-gram or neural (supervised or unsupervised) and perform reference-based quality estimation, making their performance heavily dependent on the availability and quality of these references. In contrast, reference-free metrics predict the quality of MT outputs using source-language segments without relying on gold-standard human translations [78], making them particularly suitable for low-resource settings. A few notable examples include BARTScore [79], PRISMsrc [80], and COMET-QE [81].

Although automatic metrics offer efficiency and objectivity, they may not always fully capture the multifaceted nature of translation quality, particularly in terms of contextual suitability, stylistic fidelity, and pragmatic meaning [57,82]. Moreover, their scores require careful interpretation, as their judgement depends on the design of the metric—whether it is semantically aware, has access to contextual information, or performs only lexical comparison. Recent research has therefore begun exploring evaluation approaches grounded in error annotation and error span detection mechanisms [83,84], as well as employing LLMs for error identification and interpretation [85,86]. Nonetheless, automatic evaluation metrics serve as general indicators of system performance, operating according to fixed, pre-defined criteria, which means different metrics may yield different scores for the same translation, and their reliability can vary across evaluation settings [87].

2.3.3. Context-Aware Evaluation

The study of context in MT evaluation focuses on incorporating document-level, paragraph-level, or inter-sentential information, rather than evaluating translations at the single-sentence level [88–90]. It is now widely accepted that both human and automatic evaluations should take context into account [21,28,91–93]. In particular, Castilho et al. [21] showed that nearly half of the translations they analysed required additional context for accurate assessment. Reflecting this shift, since 2022 the Conference on Machine Translation (WMT) has adopted document-level human evaluation for ranking systems in shared tasks [28,94]. Using platforms such as Appraise (<https://github.com/cfedermann/Appraise> (accessed on 20 June 2026)), human evaluators can assess either multi-sentence segments or individual sentences while having access to the full document context. However, despite these advances in human evaluation, adapting contextual principles to automatic evaluation remains challenging, and most existing automatic metrics still operate primarily at the word or sentence level.

Several attempts have been made to extend traditional AEMs to context-aware or discourse-level evaluation [89]. For example, Jiang et al. [18] introduced BlonDe, a document-level metric that incorporates features such as named entities, tense, pronouns, and discourse markers to assess translation quality beyond the sentence level. Building on this idea, Vernikos et al. [19] introduced a straightforward method to extend sentence-level metrics to the document level by taking additional sentences from the test set as context. They demonstrated improved correlation scores with human judgments by adapting three widely used metrics—COMET, BERTScore, and Prism—and COMET-QE (a reference-free metric). Their approach involved concatenating each sentence with its two preceding sentences, which were then used for embedding the reference, hypothesis, and source. Although the metric extends evaluation to the document level, it still scores sentences individually, with additional context only used during the embedding process. In the reference-free setting, Raunak et al. [20] introduced SLIDE (SLiding Document Evaluator) that operates on sentence blocks using a sliding window over the test set and accumulating scores for the entire chunk. The authors reported higher pairwise system ranking accuracy compared to sentence-level baselines (COMET-Kiwi [81]) and comparable performance with reference-based metrics, even slightly surpassing Doc-COMET [19]. For paragraph-level evaluation, Deutsch et al. [88] explored using a sentence-level dataset with concatenated multiple sentences to train a neural metric on longer texts, reporting only marginal or no improvements over baselines.

LLMs typically operate with very large context windows and, together with their in-context learning and reasoning abilities, have attracted growing interest for automatic evaluation tasks [89,95]. LLM-based metrics generally rely on prompt-based methods that instruct models to assign quality scores or to identify and categorise translation errors [85,96]. By providing multiple examples within prompts, these models can learn and contextualise their scoring strategies through in-context learning [97]. Overall, their ability to handle extended context and multilingual input highlights the potential of LLMs to reshape MT evaluation. However, several challenges remain before LLM-based metrics can be reliably adopted for primary evaluation. These include reproducibility concerns associated with proprietary models [89], inconsistent adherence to task instructions [84], potential knowledge dilution in long prompts [98], and relatively weak correlations with human judgments at the segment level [84,87].

3. Current Approaches to Automatic MMT Evaluation and Challenges

As mentioned in Section 2.3, due to high temporal and monetary costs, automatic evaluation is often the only practical way to provide feedback during model development and to track progress in the field. However, the automatic evaluation of MMT systems is inherently more challenging than that of conventional MT systems. Since most MMT models are extensions of NMT architectures, they inherit common translation issues such as grammatical errors, inappropriate lexical choices, and semantic inadequacies [99,100], while simultaneously introducing additional complexities arising from the integration of visual information [14].

In practice, evaluation continues to rely heavily on text-based AEMs and relatively short and simple test sets derived from Multi30K. These metrics and datasets provide only a narrow view of model capabilities, particularly with respect to the multimodal dimension [101,102]. Despite well-documented limitations of surface-level metrics such as BLEU, many MMT studies still report performance based solely on these scores, often without human validation or statistical significance testing [103,104]. Even advanced neural AEMs, originally designed for traditional MT systems, are commonly applied without adaptation to non-textual modalities. Consequently, such metrics may struggle to

distinguish between systems that genuinely leverage visual information for disambiguation and those that produce fluent translations independent of image input [105].

For example, Elliott [37], using METEOR text-similarity evaluation, showed that some MMT models are insensitive to visual input and produce nearly identical translations even when provided with unrelated images, suggesting that visual features are often ignored. Similarly, Raunak et al. [106] argued that the visual modality does not lead to significant gains in translation quality; however, their comparison relied on BLEU scores. Building on such findings, Wu et al. [14] questioned the validity of reported improvements in MMT and suggested that observed gains may instead result from a regularization effect. In response, Li et al. [102] highlighted limitations of evaluation benchmarks derived from Multi30K, noting that the dataset was not originally designed with multimodal evaluation in mind and advocating the creation of more suitable MMT datasets. In a similar vein, Futeral et al. [13] proposed a multilingual and multimodal contrastive evaluation set containing ambiguous sentence pairs paired with disambiguating images, requiring models to exploit visual information to generate correct translations. Datasets such as MultiSense [107] and MultiSubs [108] further expand evaluation to cross-lingual and cross-modal ambiguity.

While these efforts represent steady progress in test sets, the automatic metrics used for evaluation still require refinement to align with the domain-specific needs of MMT. As current evaluation lacks standardised vision-aware metrics capable of accurately justifying the visual context and quantifying the quality of translation. This also highlights the need for a deeper understanding of the limitations and potential of existing automatic, particularly context-aware, approaches in the context of MMT evaluation.

Based on the discussion above, we argue that the evaluation of MMT systems is a twofold problem: (1) assessing the model's visual awareness, i.e., its sensitivity to visual context and (2) evaluating the quality and correctness of the generated translation. Therefore, evaluation metrics should capture not only linguistic fidelity but also the extent to which visual information contributes meaningfully to translation quality. This raises the question of whether existing evaluation metrics, originally developed for text-only MT systems, are suitable for MMT tasks and capable of distinguishing between text-only and multimodal MT systems. In the following sections, we examine the potential of existing pre-trained evaluation metrics (Section 4) for MMT tasks and conduct experiments (Section 5) to investigate their sensitivity to multimodal information in the MMT setting.

4. Context-Aware Evaluation of MMT Systems Using Document-Level Metrics

As discussed in Section 2.3.3, document-level metrics with extended context have shown better performance in text-only tasks over baselines in both reference-based and reference-free settings. This motivates us to explore their potential for context-aware evaluation of systems that incorporate multimodal information. However, most multimodal datasets consist of independent sentence-level instances, where each sentence represents a distinct conceptual unit [102]. As a result, the direct use of document-level metrics is not feasible, since these metrics rely on neighbouring sentences for context.

A practical alternative is to use IDs as contextual input for document-level metrics when surrounding sentences are not available. This integration can be seamlessly applied without modifying the architectures of existing metrics. Since IDs represent visual content in linguistic form, their inclusion can provide useful disambiguation signals during evaluation, particularly for context-dependent translations, following the same approach used to improve contextual awareness in MT models [39,109]. This evaluation setup investigates whether integrating IDs into AEMs improves their performance in MMT evaluation, as posed in the research question.

We hypothesise that adding visual context through IDs will yield higher correlations with human judgements than traditional text-only automatic metrics in MMT tasks. To address this question and validate the hypothesis, we conduct a series of experiments using both traditional and context-aware AEMs. We compare their performance across multiple multilingual multimodal datasets with respect to correlations with human judgments. In the sections below, we explain the proposed integration of IDs in metrics and introduce the selected document-level metrics.

4.1. Image Description as Context

Visual–language pre-training has emerged as a powerful paradigm for jointly learning representations from visual and textual data [50,110,111]. Models in this category integrate visual and linguistic information by pairing strong visual encoders with large, often frozen, language models to produce unified multimodal representations (e.g., CLIP [112] BLIP-2 [50], Flamingo [52], and Gemma-3 [113]). These architectures have achieved state-of-the-art performance across a range of multimodal tasks, including image description, object detection (OD), and object recognition (OR) [114]. Recent work has explored the use of such task-agnostic visual–language models (VLMs) for image caption evaluation, aiming to assess both the linguistic and visual dimensions of candidate captions [15–17].

In order to examine the potential of such models in MMT evaluation, we first obtained the multilingual IDs for each evaluation set (Table 1) using a VLM via prompting (see Appendix Table A1 for prompts used for each language). The available test sets contain source-side ambiguous text in English, therefore to evaluate English-to-{German, Czech, French} translation direction, we generated the corresponding IDs in the target language. Directly generating visual context (which refers to the detailed contents of an image including objects, attributes, and their relationships) in the target language helps eliminate the possible linguistic or translation ambiguity [17]. In the next stage, these IDs are incorporated as contextual information within document-level metrics to assess whether they enable AEMs to more accurately assess the translation quality and align with human judgments.

Table 1. Multimodal MT evaluation test set statistics.

Dataset	No. of Segments				Total
	en→cs	en→de	en→fr	en	
CoMMuTE	308	300	308	–	916
MLT	–	381	441	–	822
test2017	–	461	461	–	922
test2016	1000	1000	1000	–	3000
ThHumB	–	–	–	500	500
Total	1308	2142	2210	500	6160

For ID generation, we employ Gemma-3-12B-IT, a multimodal LLM from the Gemma family (<https://deepmind.google/models/gemma/> (accessed on 28 June 2026)). This choice was guided by several factors including the following: the model must be open-source, multilingual, officially support image captioning and description generation, and achieve an accuracy above 50 on standard benchmark datasets. We also considered other models such as CLIP, BLIP, and Flamingo. However, following manual qualitative inspection of the generated descriptions, Gemma-3-12B-IT was selected for its better performance (qualitative evaluation was conducted on English descriptions, with the assumption that comparable performance holds across officially supported languages). Table 2 presents illustrative examples from the CoMMuTE dataset, showing source sentences, target translations, and the corresponding automatically generated IDs.

Table 2. Example from the CoMMuTE dataset, where image captions are available in English (En), with human translations in German (De), French (Fr), and Czech (Cs). The IDs are generated using an open-source VLM model.

	En: The beam seems to be moving. De: Der Strahl scheint sich zu bewegen. Cs: Zdá se, že se paprsek hýbe. Fr: Le rayon a l'air de bouger. ID: Sunlight streams through the misty trees, illuminating a gravel path leading into a dense forest.
	En: The beam seems to be moving. De: Der Balken scheint sich zu bewegen. Cs: Zdá se, že se trám hýbe. Fr: La poutre a l'air de bouger. ID: The image shows a close-up, low-angle view of an old, weathered wooden ceiling with exposed beams and rafters, some of which are inscribed with dates and names.
	En: First we should fill up the boot. De: Zuerst müssen Sie den Stiefel füllen. Cs: Je třeba nejdřív naplnit holinku. Fr: Il faut d'abord remplir la botte. ID: A pair of blue rain boots are repurposed as planters, with one containing a small green plant and soil.
	En: First we should fill up the boot. De: Zuerst müssen Sie den Kofferraum füllen. Cs: Je třeba nejdřív naplnit kufr. Fr: Il faut d'abord remplir le coffre. ID: The image shows the rear view of a teal-colored BMW M2 with a spoiler, dual exhaust pipes, and a license plate reading "EL-214-JB".

4.2. Context Integration in Document-Level Metrics

Document-level metrics typically use surrounding source and reference sentences from the same document when evaluating a given sentence. The context is extracted using configurable parameters that determine the number of preceding sentences to include [19]. In our experiments, we extend two existing document-level metrics—Doc-BERT and Doc-COMET—modifying their context integration module to use IDs instead of surrounding sentences, adapting them to the MMT task. Both of these metrics have their sentence-level baselines and open-source up-to-date implementations and are widely applied for comparing and developing MT systems [28,67,84].

Doc-BERT is an extension of BERTScore [115] to the document-level evaluation. Originally [19], the context is drawn from the reference text and added to both the hypothesis and reference sentences; then, it is excluded when calculating the similarity score for the current segment. In our implementation, we replace this textual context with IDs, extracted from relevant images using a VLM. These textual descriptions are concatenated with the hypothesis and reference before the embeddings are computed, thereby enriching their contextual representations. For the baseline pre-trained language model, we used multilingual transformer models such as roberta-large and microsoft/deberta-xlarge-mnli, selected based on their strong correlation with human judgments (https://github.com/Tiiiger/bert_score/tree/master (accessed on 16 May 2026)).

Doc-COMET [19] extends the COMET [11] framework by integrating source and reference context directly into the encoder. In the original version, COMET incorporates preceding sentences from the source text along with the current source sentence, while for the hypothesis, the context is taken from the reference to prevent potential error propagation that could occur if the hypothesis contexts were used. Similar to sentence-level COMET, document-level extension is initialised from a pre-trained language model (RoBERTa) and fine-tuned on human judgment data. In our adaptation, we replace the textual reference context with IDs, thereby allowing visual semantic information to contribute to contextual encoding. Since IDs are available only in the target language, the source context is kept same as in the original Doc-COMET setup.

5. Experiments

In this section, we describe the experimental setup, including the datasets, multi-modal and text-only MT systems, and evaluation procedure. To ensure transparency and reproducibility, all metrics are evaluated using their default configurations, and results are reported from a single run unless otherwise stated. Also, we report results obtained using publicly available datasets, open-source metrics, and pre-trained language models, including LLMs. For trained and pre-trained models such as BERTScore and COMET, we employ the best-performing (as per their official results) publicly available models with default hyper-parameter settings. For meta-evaluation, we consider system performances, translation accuracy on contrastive test-sets, and correlation with human scores where human annotations are available.

5.1. Dataset

We use a combination of evaluation sets from the image caption and MT domains, including Multi30K, MSCOCO-test2017, MLT, CoMMuTE, and THumB. These datasets cover a range of evaluation aspects including ambiguous lexical terms, verbs, and contrastive multilingual translations, providing the means to assess both lexical accuracy and multimodal sensitivity. These datasets are described below:

THumB 1.0 [116] comprises 500 images sourced from MSCOCO [117], each paired with one human-written caption and four automatically generated captions. The evaluation of candidate captions involves manual assessment along three dimensions: precision (the extent to which the caption accurately describes the image), recall (the degree to which it captures salient information within the image), and total (an overall quality score considering fluency, inclusivity, and conciseness). To support automatic evaluation, the dataset contains 4 reference captions for each image. The presence of both human-authored captions and human ratings makes this dataset a valuable resource for meta-evaluation. However, its main limitation lies in its monolingual nature, as the image captions are evaluated in the English language only. We utilise THumB to estimate overall system performance and compute metrics' correlation with human judgements.

MLT [34] is constructed by identifying 1108 English words that are ambiguous (i.e., they have multiple possible translations in the target language). Each word appears across multiple sentences, resulting in a total of 98,647 MLT data points. The dataset, available for both English–German and English–French language pairs, is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Its multimodal design, with identified ambiguous lexical items, can be helpful in studying visual disambiguation capabilities of models. In this study, we use MLT to investigate whether systems with stronger visual-awareness receive correspondingly higher ratings from AEMs compared to image-agnostic systems.

test2017 (MSCOCO) [8] is a standard evaluation set for the MMT task, comprising IDs that often contain ambiguous verbs. The dataset is derived from images sampled from the VerSe dataset [118], which in turn was adapted from the MSCOCO dataset [117]. Ambiguous verbs with multiple senses were first identified in VerSe using the IDs in English, and their corresponding translations were then collected in the German and French language. Owing to its design and relevance to lexical ambiguity, MSCOCO-test2017 has become a widely adopted benchmark for evaluating MMT systems [8].

test2016 (Multi30K) is another widely used benchmark for evaluating MMT systems. It is a test split of the Multi30K dataset [36], which itself is an extension of the Flickr30K dataset [119]. The original Multi30K corpus comprises 31,014 images paired with English descriptions translated into multiple languages. While the IDs were collected via Amazon Mechanical Turk, the translations were produced by professional native translators, ensuring high linguistic quality and cross-lingual consistency. This test-set comprises 1000 English segments and their translations in Czech, German, and French.

CoMMuTE (Contrastive Multilingual Multimodal Translation Evaluation) [13] contains 155 lexically ambiguous English sentences, each paired with two plausible translations in the target language. Typically, each translation is correct given the accompanying image, which serves to disambiguate the intended sense of the ambiguous source word. The ambiguous items are either drawn from the discourse sets of Bawden et al. [120] or newly introduced by the authors. Images were sourced online and chosen specifically to resolve the ambiguity in one of the two senses. The dataset includes English–French, English–German, and English–Czech language pairs, providing exactly two translations and one disambiguating image per instance. By offering contrastive multilingual translations, CoMMuTE supports and enables the examination of metric robustness in scenarios with noisy or suboptimal references.

5.2. Translation Systems

The evaluation datasets (Section 5.1) are provided as triplets comprising the source text, reference translation, and associated image. However, they do not include hypothesis translations—an essential component for automatic evaluation. To address this, we generate hypotheses using a diverse set of translation systems, including NMT-, MMT-, and LLM-based models, spanning both text-only and multimodal configurations. The translation systems used in this study are described in detail below.

ZeroMMT [38] is a zero-shot MMT approach that bypasses the need for large-scale multimodal parallel data by adapting a strong text-only MT model with lightweight visual adapters during training. Given the scarcity of multimodal data, MMT systems typically underperform compared to standard NMT models. To address this, the authors extend strong baseline NMT models with multimodal capabilities, removing the need of costly supervised data. The approach leverages visually conditioned masked language modelling (VLM) to integrate visual context into translation and employs Kullback–Leibler (KL) divergence to align outputs between multimodal and text-only MT systems. Validation was performed on multiple languages, including German, Russian, French, Czech, Arabic, and Chinese, using BLEU and COMET. In our experiments, ZeroMMT represents purpose-built strong MMT to facilitate comparison with traditional MT systems.

Opus-MT [121] is a collection of open-source bilingual and multilingual MT models, primarily trained on the OPUS parallel corpus. The project aims to provide freely accessible translation services and tools, independent of commercial interests or restrictions. OPUS-MT hosts over 1000 pre-trained models (<https://huggingface.co/Helsinki-NLP> (accessed 8 July 2026)) covering a wide range of language pairs. The models are based on the standard Transformer architecture and are implemented using the Marian-NMT framework

(<https://marian-nmt.github.io/> (accessed 24 May 2026)). For our experiments, we use the bilingual models available on Hugging-Face and take Opus-MT as a representative of open-source bilingual text-only MT systems.

Aya-23-8B [122] is a multilingual LLM based on an optimised Transformer architecture and fine-tuned to follow human instructions. Aya-23-8B is a text-to-text model, accepting only text inputs and producing text outputs. The model supports 23 languages and contains 8 billion open-source parameters. Although Aya-23 was primarily trained on general text data and not specifically fine-tuned or evaluated for MT tasks, it demonstrated strong performance on the WMT24 translation benchmarks [28]. Many recent LLMs exhibit inherent multilingual and translation capabilities, yet there is no systematic and reliable evaluation of their MT performance. For this reason, we include a text-only LLM (Aya-23-8B) in our experiments on MMT evaluation. Model selection was guided by reported language coverage in the original documentation, ensuring that both source and target languages of our test sets are supported. We further prioritised open-source models to ensure reproducibility of results. The prompt used for translation is given below:

“Translate the following sentence into {German | French | Czech}, output only the translation, nothing else.”

Gemma-3-12B-IT [113] is an instruction-tuned multimodal LLM. Instruction tuning enables the model to follow task-specific prompts and natural language instructions, making it well suited for our task where multimodal context can be added via prompting. Gemma-3-12B-IT supports long context windows of up to 128K tokens and offers multilingual capabilities. Its visual grounding can help address lexical ambiguities and context-dependent meanings that are challenging for text-only MT systems. We employed Gemma-3-12B-IT for multimodal translation via prompting, instructing the model to translate the source text while considering the associated image, with the following prompt:

“Translate the following sentence into {German | French | Czech}, using the image as context. Output only the translation, nothing else.”

5.3. Baseline Metrics

We use several baseline metrics to compare our proposed methods against well-studied approaches commonly used to report progress in MMT research, we describe those baselines below.

BLEU [9] is a widely used metric for MT evaluation that measures the precision of n-grams between candidate and reference translations. Its popularity stems from computational efficiency and its ability to capture local word-level correspondence. However, BLEU is criticised for notable reasons: it disregards grammatical correctness, tends to favour shorter translations, and struggles to account for valid paraphrases or synonymous expressions that preserve meaning [71,103]. Despite these limitations, BLEU continues to be frequently used to report and interpret progress in recent MT research [38,84], mainly due to its language independence, computational simplicity, and long-standing acceptance within the research community.

TER (Translation Edit Rate) and its variant Human TER [64] quantify the number of edits required to transform a machine-translated output into a reference translation. By focusing on the post-editing effort, TER provides a direct and interpretable measure of the changes necessary to produce an acceptable translation. This emphasis on edit operations makes TER particularly useful for assessing practical translation quality and post-editing effort.

CHRF [68] is a character n-gram F-score metric that evaluates translations based on character-level matches rather than word-level correspondence. This design makes CHRF particularly effective for morphologically rich languages, as it is largely language-independent

and does not require tokenization. As a result, CHRF has been shown to correlate well with human judgments—often outperforming more complex word-level metrics—and remains a reliable choice, especially for low-resource languages [67].

BERTScore [115] is an unsupervised metric for evaluating text generation that leverages contextual embeddings for soft comparison between reference and candidate translations. Tokens from both the reference and hypothesis are represented using embeddings from models such as BERT or RoBERTa, and precision, recall, and F1 scores are computed based on the best-matching token embeddings. BERTScore has been shown to outperform traditional untrained metrics, demonstrating stronger alignment with human judgments [73]. In our experiments, we only report F1 scores with TF-IDF weighting in our experiments and adapted the models from official implementation (https://github.com/Tiiiger/bert_score/tree/master (accessed 18 March 2026)).

COMET [11] is a supervised neural evaluation metric that differs fundamentally from string-based metrics (such as BLEU, TER, and CHRF) and embedding-based approaches such as BERTScore. The COMET framework includes both reference-based and reference-free variants, trained on human judgment data such as Direct Assessments (DAs) [58], HTER [64], and Multidimensional Quality Metrics (MQMs) [60] and supports a wide range of languages. In our experiments, we use the default `wmt22-comet-da` model available on HuggingFace (<https://huggingface.co/Unbabel/wmt22-comet-da> (accessed 18 March 2026)), which is initialised with XLM-RoBERTa as its backbone. This model computes scores using the source text, hypothesis translation, and reference translation, enabling it to capture semantic similarity and meaning preservation more effectively than surface-form metrics [123].

6. Meta-Evaluation

There are multiple approaches to metric meta-evaluation, which aims to quantify how well automatic metrics agree with human judgments of translation quality. Agreement can be measured using ranking-based approaches, such as Spearman's ρ or Kendall's τ , or using linear correlations, such as Pearson's coefficient, each capturing different aspects of metric performance [124]. Additionally, metric evaluation can be defined at various levels, including segment-level or system-level, and can focus on specific language pairs or domains.

One of the primary use cases of automatic evaluation metrics is ranking MT systems during development or when comparing systems from different competitors. Such rankings can be useful for shortlisting or clustering systems based on performance. For example, in WMT shared tasks, the large number of participating systems are shortlisted using AEMs for subsequent manual filtering [67]. Ideally, AEMs should rank MT systems in a manner consistent with human judgments, and their effectiveness is evaluated based on how closely their rankings align with those of humans [67,125]. However, in our study, the number of systems is limited (four in total), which does not provide sufficient statistical power for pairwise system-level accuracy or significance testing between system rankings [126]. Therefore, we rely on traditional correlation-based measures and accuracy on contrastive evaluation for meta-evaluation.

In the following sections, we first present the overall translation performance of the MT and MMT systems (Section 6.1). This is followed by an analysis of the metrics' correlation with human judgment (Section 6.2) and an assessment of the contextual awareness of the document-level metrics (Section 6.3). Finally, we report the results from a small-scale human evaluation (Section 6.4) to provide qualitative insights.

6.1. System-Level Results

This section provides an overview of the performance of the selected MT systems across different metrics on the evaluation set. System scores are computed as the average of

their segment-level performance. These averages can be further aggregated across systems or languages to obtain an overall score per metric or per system.

It is important to note that these results are not intended for direct comparison between metrics or MT systems. Rather, they illustrate the score allocation strategies of different metrics and provide context for the meta-evaluation presented in subsequent sections. Additionally, these results highlight how the magnitude of scores can be different for the same systems when evaluated by distinct metrics, as different AEMs employ different scoring strategies, reflecting their design and evaluation objectives. Some metrics adopt a more rigid approach (exact matches), while others allow for softer comparisons (embedding-based) between reference and candidate texts.

Table 3 presents the system-level scores, computed by aggregating scores across metrics and languages per system. Bold values in each row indicate the system with the highest average score for a given metric. Overall, the scores suggest that systems trained on multimodal data (ZeroMMT and Gemma-3) outperform text-only systems. Among the text-only systems, Aya-23 and Opus-MT achieve comparable scores but consistently underperform relative to the multimodal systems.

Table 3. System-level scores computed by averaging segment-level scores assigned by AEMs across evaluation corpora. For each metric, the highest score is shown in bold, together with the difference from the lowest score in that row. An upward arrow (↑) indicates that higher values correspond to better performance, whereas a downward arrow (↓) indicates that lower values are better.

Metric	Aya-23	Gemma-3	Opus-MT	ZeroMT
BLEU	34.50	36.68	36.19	38.39 (+3.89 ↑)
CHRF	59.59	61.69	60.35	61.85 (+2.26 ↑)
TER	48.48	46.75	46.46	44.93 (−3.55 ↓)
BERTScore	0.693	0.707	0.702	0.716 (+0.02 ↑)
Doc-BERT	0.734	0.750	0.740	0.755 (+0.02 ↑)
COMET	0.815	0.836 (+0.04 ↑)	0.800	0.831
Doc-COMET	0.732	0.763 (+0.05 ↑)	0.718	0.753

According to the document-level extension of COMET (i.e., Doc-COMET), the largest score difference is 0.05 between Gemma-3 and Opus-MT. Standard COMET reports a similar difference of 0.04 points between the same systems (see the last two rows of Table 3). The next highest score difference, 3.89, is observed in BLEU for ZeroMMT over Aya-23. Notably, Aya-23, trained on 8 billion parameters, is still outperformed by ZeroMMT, which adapts a 1.3 billion parameter NLLB model [127] through visually conditioned language masking. Appendix C provides a more detailed comparison of metric scores according to language, system, and corpora.

Overall system-level scores support the Multimodal MT systems, for example, both embedding-based metrics and string-based metrics consistently rank ZeroMMT higher, while supervised metrics favour Gemma-3, the multimodal LLM, over standard MMT (i.e., ZeroMMT). This difference in metric agreement likely arises from two factors: first, embedding-based and string-based metrics tend to favour lexically simpler translations, as seen in ZeroMMT, and may penalise the more varied vocabulary produced by Gemma-3. Second, trained neural metrics such as COMET and Doc-COMET, which are fine-tuned on human annotations, better capture linguistic variation and align with human preference, as discussed in Section 6.4.

6.2. Correlation Scores on Image Caption Test-Set

One of the ways to measure the performance of automatic evaluation metrics is to find how well they correlate with human judgments [69,128]. For this, we use the

THumB dataset, which collects human ratings for automatic image captioning (IC) along three dimensions: precision, recall, and overall quality. We focus on the overall quality score. Each image in THumB includes four candidate captions from different automatic captioning systems and five human-written references. For the overall performance overview of AEMs, one human reference is treated as a candidate caption and was excluded from the references during evaluation. Moreover, there are multiple correlation measures, such as Pearson, Spearman, and Kendall, each with its own score range, strengths, and limitations [129]. To ensure a fair comparison, we compute all three correlation coefficients to examine the performance and suitability of existing AEMs in the context of MMT evaluation.

Figure 2 presents the correlation scores on the THumB dataset. Overall, COMET shows the strongest alignment with human judgments at both the sentence and document levels, with Pearson’s correlations around 0.31–0.33, Spearman around 0.26, and Kendall around 0.19. Although the differences between Doc-COMET and COMET are not statistically significant, they are consistent across all correlation types. COMET considers both the source text and the reference when scoring MT outputs, while Doc-COMET additionally incorporates IDs as context, which likely helps produce automatic scores closer to human ratings.

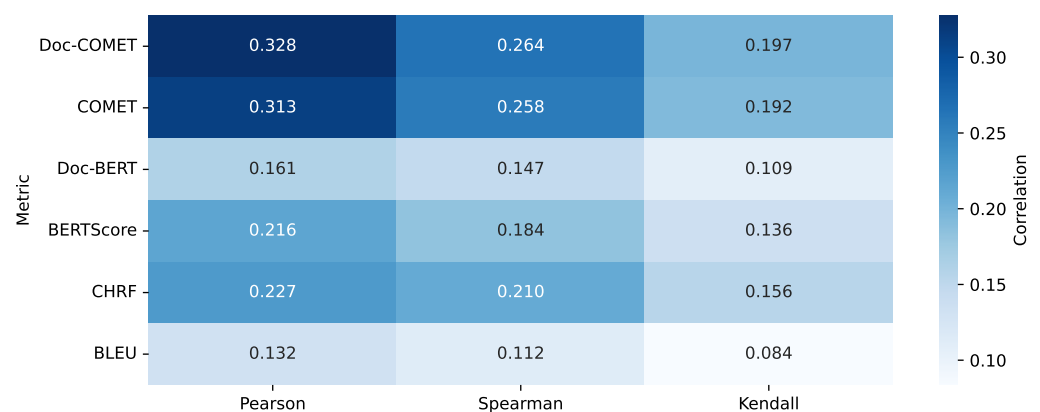


Figure 2. Heat-map of Pearson, Spearman, and Kendall correlations between human evaluation scores and automatic metrics (COMET, Doc-COMET, BERTScore, Doc-BERT, and BLEU). Overall COMET exhibits the highest correlations across all measures, while BLEU shows the weakest alignment with human judgments.

Notably, the character n-gram-based metric CHRF achieves the second-highest correlation scores, outperforming embedding-based metrics (BERTScore and Doc-BERT) and showing the continued value of lexical overlap-based evaluation. The performance of CHRF aligns with recent findings from the WMT2025 Metric Shared Task, where CHRF achieved results comparable to the strongest neural submissions [84]. Surprisingly, Doc-BERT, despite incorporating additional IDs, performs worse than its sentence-level counterpart. Since BERTScore measures soft semantic similarity between embeddings, the increased token length introduced by additional context may dilute the similarity signal, making it harder to penalise semantically divergent yet fluent translations [74]. BLEU exhibits substantially lower correlations (Pearson’s ≈ 0.13 , Spearman ≈ 0.11 , and Kendall ≈ 0.08), reflecting its limited alignment with human judgments.

Graham and Baldwin [130] argue that higher correlations with human judgments may arise by chance and therefore recommend statistical significance testing to determine whether differences between the correlations of two metrics are meaningful. We therefore apply the Williams test for statistical significance. Figure A2a shows Pearson’s correlation between all metric pairs, where almost all correlations are high, with several exceeding $r = 0.9$. The Williams significance test results are shown in Figure A2b. Contrary to what correlation rankings might suggest, Doc-COMET—despite achieving the highest Pearson’s

correlation with human judgements (Table 2)—is not significantly better than any other metric. Conversely, CHRF, which correlates less strongly with human judgements than COMET and Doc-COMET, is shown to be significantly better than both. Although these findings highlight the limitations of relying solely on correlation as an indicator of metric quality, it is worth noting that correlation differences among metrics are moderate, which makes the significance test results unsurprising.

Taken together, these results suggest that learned neural metrics provide a closer approximation to human evaluation on multimodal translation tasks, consistent with their performance in NMT [69]. Incorporating IDs slightly improves performance, highlighting the value of contextual information. However, additional context may not benefit all metrics; for example, BERTScore performance decreases after integrating context. Furthermore, the systematic decline in correlation values across metrics underscores the stricter nature of rank-based measures (e.g., Spearman; Kendall) and emphasises the importance of reporting multiple correlation types to obtain a more comprehensive assessment of metric reliability [131].

6.3. Contrastive Evaluation

Contrastive or challenge sets are commonly used for targeted evaluation of NLP systems, enabling focused assessment of specific linguistic phenomena or revealing system strengths and weaknesses that are not apparent in standard test sets [84]. Unlike conventional evaluation corpora, which typically contain generic, real-world text, challenge sets include rare, adversarial, or ambiguity-focused examples. For instance, the CoMMuTE test-set [13] provides contrastive translation pairs accompanied by visual context: an ambiguous English source sentence, its correct human translation, and an incorrect (*bad_reference*) translation that intentionally mistranslates the ambiguous word.

To assess the robustness of metrics under adversarial conditions, we conducted a contrastive evaluation on the CoMMuTE dataset. We performed two evaluation runs: one using the correct reference and another using the *bad_reference*, which introduces a controlled semantic error (see Figure 1). A robust evaluation metric should assign a lower score to a system output when compared against the incorrect reference than when compared against the correct one. This setup enables us to assess not only the sensitivity of automatic metrics to subtle meaning shifts but also whether the MT systems themselves successfully resolve lexical ambiguities, as reflected by differences in metric scores across the two contrastive runs.

Figure 3 presents a heatmap of the average score differences for MT systems. The visualisation clearly separates text-only from multimodal systems. For the text-only systems (Opus-MT, a bilingual MT model, and Aya-23, an LLM-based MT system), most automatic metrics show almost no score differences between the *good* and *bad_reference* variations. In contrast, systems that incorporate visual information (ZeroMMT and Gemma-3) exhibit more pronounced differences, underscoring the usefulness of visual context for producing translations different from traditional systems. Results on the contrastive set further reinforce the overall system-level findings (Table 3), with multimodal MT systems achieving higher average scores than their text-only counterparts.

At the metric level, Doc-COMET demonstrates the highest sensitivity to quality variations, yielding slightly larger score differences than all other metrics—specifically, 2.9% more than the baseline COMET. N-gram-based metrics perform comparably to neural metrics in detecting lexical differences for multimodal systems. BLEU and CHRF, being precision-oriented metrics, tend to perform well on contrastive sets containing simple and repetitive sentences, where differences often hinge on subtle lexical changes [132]. Table 4 shows the computed correlations between the *good* and *bad_reference* scenario, where a

higher correlation coefficient indicates a metric’s inability to detect noise and vice versa. Metrics such as COMET ($r = -0.14$) and Doc-COMET ($r = -0.42$) show strong sensitivity to corrupted references, producing low or negative correlations that suggest they sharply penalise mistranslations when reference quality declines. Similarly, chrF ($r = 0.45$) and BERT-based metrics ($r \approx 0.41$ – 0.47) demonstrate moderate sensitivity, while BLEU ($r = 0.62$) and TER ($r = 0.84$) preserve relatively higher correlations, indicating weaker responsiveness to reference errors. In this framing, low correlations are not evidence of metric unreliability but rather a sign of robustness, as they demonstrate the metric’s capacity to detect and react to poor reference quality rather than smoothing over it.

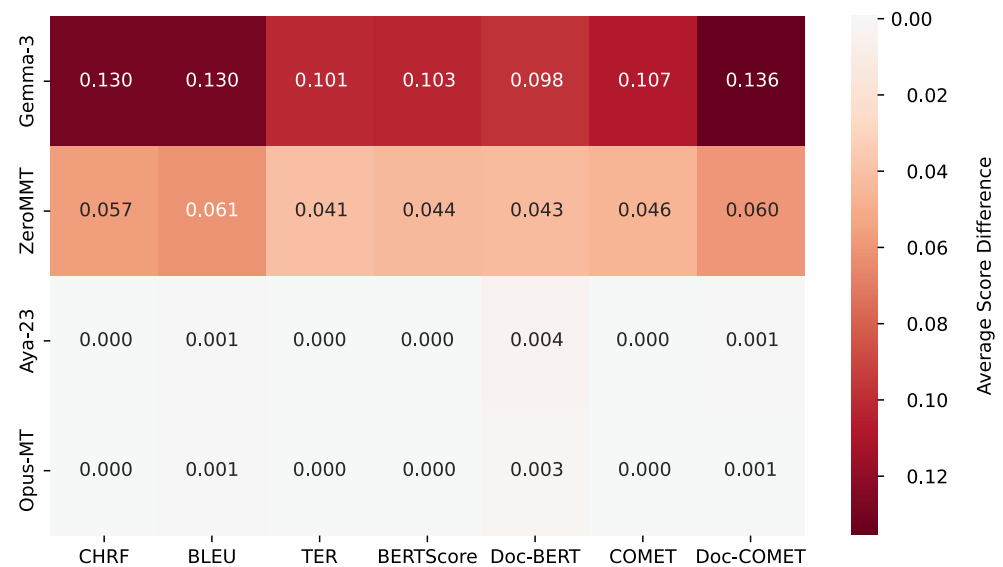


Figure 3. Heatmap of average score differences across different MT systems and automatic metrics (COMET, COMET doc-level, BERTScore, BERT doc-level, and BLEU). Doc-COMET shows the largest differences, slightly higher than those of n-gram metrics. Interestingly, text-based MT systems exhibit almost zero sensitivity to bad references.

Table 4. Original and degraded (bad_reference) Pearson correlation coefficients (r) and averaged metric scores obtained across systems on the CoMMuTE test set.

Metric	Score (Original)			Score (Bad_Reference)			r
	cs	de	fr	cs	de	fr	
BLEU	0.31	0.39	0.41	0.27	0.35	0.36	0.622
CHRF	0.51	0.57	0.60	0.47	0.53	0.55	0.454
TER	0.44	0.53	0.54	0.41	0.50	0.50	0.843
BERTScore	0.68	0.71	0.74	0.64	0.67	0.70	0.477
Doc-BERT	0.71	0.74	0.76	0.67	0.70	0.72	0.410
COMET	0.84	0.81	0.81	0.81	0.78	0.77	−0.143
Doc-COMET	0.75	0.73	0.72	0.71	0.68	0.67	−0.421

In summary, both surface-level and learned metrics were able to distinguish translation variations in the contrastive evaluation. Doc-COMET, having IDs as context and source translation in consideration, shows better suitability for multimodal tasks. Overall COMET, as a task-specific trained metric, demonstrated higher sensitivity compared to BERTScore, which relies on general-purpose pre-trained embeddings.

6.4. Human Evaluation Results

To analyse how visual context affects human perception of translation quality, a human evaluation study was conducted on a subset of the CoMMuTE testset using the Appraise platform. The human evaluator rated translations on a 0–100 scale under two conditions:

text-only (source and hypothesis only) and multimodal (source, hypothesis, and image). The interface for multimodal evaluation is shown in Appendix B. Each instance was evaluated multiple times to obtain robust averages. The evaluator was a bilingual German–English speaker with German as their native language and an educational background in computational linguistics.

Table 5 presents the average score assigned by the human evaluator and the automatic metrics, with the correlation score in the last column. For the without-image (w/o img) condition, all systems received an average human score above 90 on the 0–100 scale for English→German translation. This skewed average is obvious for two reasons. First, the MT systems—including the LLM-based models (Aya-23 and Gemma-3)—were trained on large, high-quality German corpora and were tested on relatively simple image-caption translations. Second, without access to the image, the human annotator cannot judge whether the system resolved the lexical ambiguity correctly, and therefore tended to assign high scores to translations that were lexically and syntactically well formed. However, when the image context was available, the average score drops, indicating that the human evaluator revised the ratings.

Table 5. Human and automatic scores computed on a subset of the COMMUTe corpus (100 segments per system), with Pearson correlation coefficients (r) shown in the final column. “Img” and “w/o Img” denote whether human assessments were conducted with image context or without image context, respectively.

Metric	Aya-23	Gemma-3	Opus-MT	ZeroMT	r
Human (Img/w/o Img)	45.0/100.0	93.0/96.2	48.6/97.2	74.2/96.8	–
CHRF	45.40	58.33	51.25	55.26	0.923
BLEU	26.36	36.15	31.79	34.24	0.881
BERTScore	0.61	0.72	0.66	0.68	0.923
Doc-BERT	0.65	0.75	0.69	0.71	0.945
COMET	0.73	0.83	0.77	0.79	0.951
Doc-COMET	0.61	0.74	0.65	0.69	0.982

This scoring strategy is visualised in Figure 4, which presents the segment-level score distributions for the human evaluator and the automatic metrics. The diagonal plots show the score distributions of individual metrics, while the off-diagonal plots illustrate pairwise comparisons between metrics. The distributions reveal clear shifts in both human and automatic scores depending on the availability of an image during evaluation. The score distribution for the human evaluator confirms that for text-only conditions, the human evaluator tended to assign relatively high scores, particularly to fluent translations with plausible semantics. When the image was available, the score distribution became more dispersed, with pronounced peaks at both 0 and 100. These findings reinforce that even small mistranslations or subtle lexical errors can have a substantial impact on perceived translation quality when expert evaluators have access to visual context, in line with previous observations by Lala [12].

Among the AEMs, COMET and Doc-COMET exhibit similar bimodal distributions, following the human scoring behaviour. This pattern reflects cases where MT systems failed to resolve lexical ambiguities correctly, for example, Aya-23 and Opus-MT received scores at both extremes depending on whether ambiguity was resolved or not, as shown in Table 5. String-based metrics such as BLEU and CHRF show a comparable bimodal trend, distinguishing between text-only and multimodal MT outputs. However, due to their reliance on exact surface-form matching and sensitivity to lexical variation, these metrics produce lower absolute score ranges than neural metrics, resulting in larger peaks at the lower end of the scale.



Figure 4. Score distribution showing how different metrics and humans have assigned segment-level scores for English–German. Graphs along the diagonal show distributions of each metric, while the off-diagonal represent comparisons between metrics’ scores, with each point representing a segment. Note that the score range differs across metrics.

In contrast, embedding-based metrics such as BERTScore and Doc-BERT assign more generous scores, leading to more normally distributed outputs. Because these metrics rely on cosine similarity between contextual embeddings rather than strict n-gram overlap, they may overestimate semantic similarity in some cases. This makes them vulnerable to structurally similar but semantically divergent translations (e.g., “man bites dog” vs. “dog bites man”), as previously highlighted by Kaster et al. [133].

Among the systems, Gemma-3 (a multimodal LLM) and ZeroMMT (an MMT system) achieved the highest and second-highest human scores (Table 5), respectively. In contrast, Aya-23 and Opus-MT exhibited substantial drops in human scores when evaluated with the associated image, indicating that these text-only systems often failed to resolve visual ambiguities that became apparent once the image was visible.

In summary, the human study confirms that context plays a critical role for both human evaluators and automatic metrics, as both human and some AEMs remained sensitive to visual context. The annotator substantially revised their judgments once the image was available, demonstrating that text-only evaluation can mask important differences between multimodal and text-only MT systems. Likewise, trained metrics—particularly learned neural metrics—show improved performance and stronger correlations with human scores when visual information is incorporated via IDs.

It is important to note, however, that the human evaluation was based on source-candidate comparison, whereas AEMs performed reference-based evaluation, with COMET being an exception in that it also considers the source text. Since each metric captures

different aspects of translation quality (e.g., lexical overlap, semantic similarity, and syntactic structure), and lexical matching remains a common foundational signal across metrics [133]; combining multiple metrics or reporting them jointly can provide a more comprehensive and reliable assessment.

7. Discussion and Conclusions

This study examined how well existing AEMs perform for the MMT evaluation task and whether incorporating IDs as context improves their reliability. We evaluated a range of metrics (including n-gram-based, embedding-based, and supervised neural metrics) across multiple test sets, comparing their behaviour on translations produced by text-only and multimodal MT systems.

Across a series of experiments on multimodal evaluation sets, supervised neural metrics demonstrated better performance. In particular, Doc-COMET, which incorporates visual semantics extracted from associated images, consistently outperformed both string-based and embedding-based baseline models. At the system level, Doc-COMET assigned the highest scores to Gemma-3, a multimodal LLM-based system, with a clear margin over text-only baselines. This outcome was also reflected in the human evaluation results (Table 5). The performance of Doc-COMET on both contrastive evaluation and correlation with human judgments highlights the importance of contextual information for assessing MMT output. Notably, Doc-COMET's evaluation strategy resembles human assessment practices, as it considers the source text, the candidate translation, reference translations, and image-derived semantic information. These supervised metrics achieved the highest agreement with human ratings across all correlation measures.

In response to the research question (whether adding visual context in the form of IDs benefits neural metrics during automatic evaluation), the results indicate moderate but consistent improvements over context-agnostic counterparts, suggesting that contextual information can improve the reliability of metric predictions. However, adding visual context does not guarantee better performance. The effectiveness of contextual integration depends on the metric architecture. Doc-BERT, for example, did not outperform BERTScore despite using additional context, which contrasts with earlier findings [19,89]. In most cases, Doc-BERT performed worse than its sentence-level variant, suggesting that the longer input sequences introduced by IDs may dilute semantic signals rather than strengthen them. Moreover, both BERT-based metrics underperformed compared to n-gram-based metrics, particularly in cases where incorrect translations remained lexically similar to the reference [74].

Overall, the performance gap between n-gram metrics and learned metrics (supervised or embedding-based) was small, and all metrics showed positive correlation with human ratings. Yet the contrastive and human evaluation results on the CoMMuTE corpus revealed clear differences in how metrics respond to visually grounded meaning. All metrics except COMET and Doc-COMET showed positive correlation with `bad_reference`, indicating limited ability to detect small but meaningful semantic changes that humans heavily penalised. In contrast, Doc-COMET showed the strongest ability to capture quality degradation under corrupted references. While n-gram metrics (BLEU, TER, and CHRF) handled lexical mismatches reasonably well—sometimes better than embedding-based metrics in short-caption data—their reliance on surface overlap limits their ability to detect deeper semantic errors. The human evaluation further confirmed the importance of visual context: without images, the annotator assigned near-ceiling scores to all systems, but once images were shown, scores changed sharply, especially for text-only systems. This shows that textual fluency alone can mask semantic inaccuracies that become clear only when visual grounding is considered. Multimodal systems achieved the highest human scores, and trained metrics that used IDs showed better alignment with these expert judgments.

In summary, three main conclusions emerge:

- First, visual context plays a meaningful role in both translation generation and evaluation, particularly for content involving ambiguity or visually grounded meaning. Models incorporating visual information (Gemma-3 and ZeroMT) achieved higher average human ratings, and the inclusion of image context led to improved alignment between metric scores and human judgments.
- Second, learned metrics such as COMET and Doc-COMET are better suited than embedding-based metrics such as BERTScore and Doc-BERT for evaluating MMT quality, as they reflect human preferences more closely and better distinguish between multimodal and text-only systems.
- Third, the mixed behaviour of different metrics highlights the need for multi-metric evaluation. Each metric captures different aspects of translation quality, and no single metric is sufficient for a comprehensive assessment. Although learned metrics perform well, n-gram metrics remain essential for low-resource scenarios where language-independent tools are required [67]. Simple parameter-free averaging of multiple metrics has also been shown to improve correlation with human judgments by up to 13% [133].

While the findings support the usefulness of IDs in multimodal evaluation, this study also reveals limitations. The improvements from contextual integration remain moderate, and not all metric architectures benefit from additional inputs. Moreover, the human evaluation was small in scale, limiting broader generalisation. Future work should expand human evaluation efforts and explore more effective ways of integrating visual grounding directly into metric design, beyond the shallow use of descriptive context adopted here. One possible direction is to compute embedding similarity using multimodal-multilingual models (e.g., ViLBERT; CLIP) instead of unimodal BERT, as demonstrated in related work on image caption evaluation [15,16]. Another avenue is to initialise evaluation metrics from vision–language models rather than text-only language models, followed by fine-tuning on human evaluation data, to test whether multimodal pre-training helps reduce token bias and ambiguity in metric predictions.

Author Contributions: Conceptualization, S.U.H., S.C. and Y.G.; methodology, S.U.H. and Y.G.; formal analysis, S.U.H.; writing—original draft preparation, S.U.H.; writing—review and editing, S.C. and Y.G.; supervision, S.C. and Y.G.; funding acquisition, S.C. and Y.G. All authors read and approved the final manuscript.

Funding: This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. This research was also supported by Research Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT Centre at Dublin City University and Trinity College Dublin. ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology, is funded through the Research Ireland Centres Programme. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by Faculty Research Ethics Committee, Faculty of Humanities & Social Sciences, Dublin City University (protocol code: DCU-FHSS-2024-043 and date of approval: 4 July 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data and code produced during the research is freely available at following github link: <https://github.com/sami-haq99/Visual-semantic> (accessed on 15 December 2025).

Acknowledgments: We would like to express our sincere gratitude to Hoang-Bao Le and Monja Jannet for their thoughtful and constructive feedback which greatly contributed to improving the quality of our work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AEM	Automatic Evaluation Metric
MT	Machine Translation
MMT	Multimodal MT
ID	Image Description
LLM	Large Language Model

Appendix A

We provide the prompts used to generate image descriptions.

Table A1. Prompts used to generate image descriptions. Image was provided with each prompt in a json format for generation.

Language	Prompt
English	Describe the image in a single sentence in English. Provide only the description, nothing else.
French	Décrivez l'image en français en une seule phrase. Ne fournissez que la description, rien d'autre.
German	Beschreiben Sie das Bild in einem Satz auf Deutsch. Geben Sie nur die Beschreibung an, sonst nichts.
Czech	Popište obrázek jednou větou v češtině, uveďte pouze popis, žádné další informace.

Appendix B

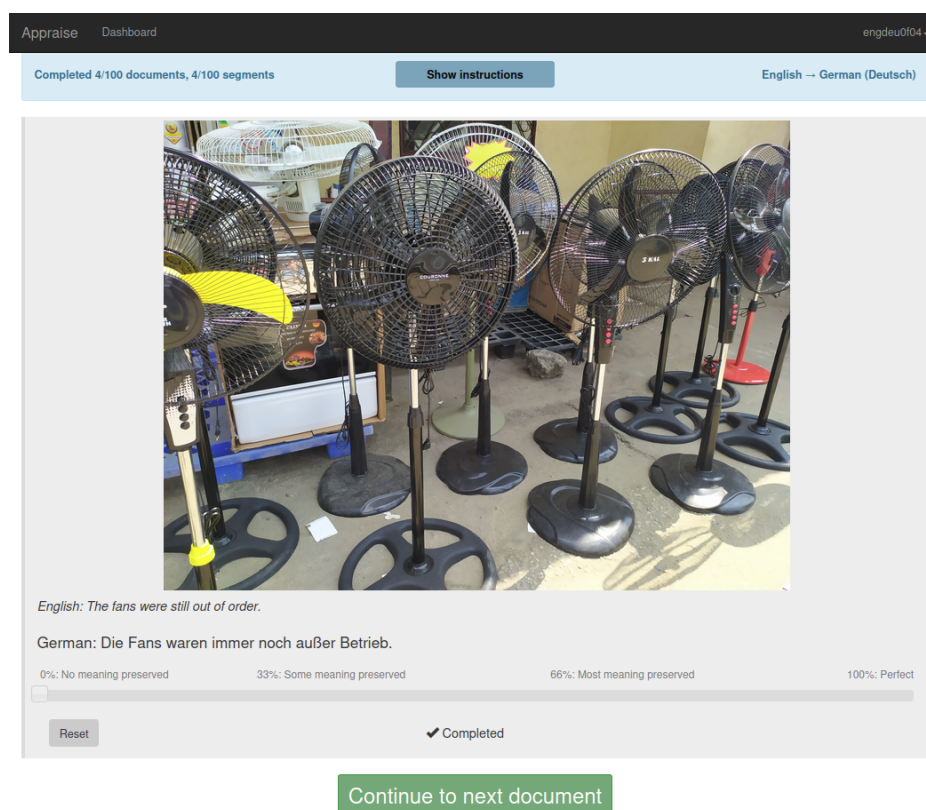


Figure A1. Screenshot of the Appraise interface showing the English source, the German translation, and the associated image as additional context.

Appendix C

The table below presents a comprehensive breakdown of AEMs across the evaluation corpora, reporting results for three language pairs (en-cs, en-de, and en-fr) as well as averages of corpus scores and language-pair scores per metric. Bold values in the corpus section indicate the highest average score for a given language pair, while bold values in the last column represent the highest average score across corpora. For all metrics, the test2016 and test2017 corpora achieve the highest average scores, with test2016 topping five out of seven cases; these test sets also contain the largest number of segments (see Section 5.1). Similarly, for language pairs, en-fr achieves the highest average score in most cases, with en-cs being the top scorer in two out of seven instances.

Table A2. Evaluation scores from automatic metrics across four corpora and three language pairs (en→cs, en→de, and en→fr), reporting corpus-level, metric-level, and overall average scores.

Metric	Corpus	cs	de	fr	Ave. Corpus
CHRF	CoMMuTE	51.43	57.06	60.26	56.25
	test2017	-	60.46	70.91	65.69
	test2016	57.24	66.36	71.32	64.97
	MLT	-	48.94	64.70	56.82
	Ave. language	54.34	58.21	66.80	60.93
BLEU	CoMMuTE	31.17	39.02	41.80	37.33
	test2017	-	33.01	50.96	41.98
	test2016	31.90	39.74	51.34	40.99
	MLT	-	8.70	36.77	22.74
	Ave. language	31.54	30.12	45.22	35.76
TER (↓)	CoMMuTE	55.61	47.26	45.97	49.61
	test2017	-	54.55	33.75	44.15
	test2016	50.94	44.96	33.28	43.06
	MLT	-	55.69	44.52	50.11
	Ave. language	53.25	50.62	39.38	46.73
BERTScore	CoMMuTE	0.678	0.706	0.739	0.708
	test2017	-	0.677	0.789	0.733
	test2016	0.725	0.754	0.812	0.764
	MLT	-	0.502	0.661	0.582
	Ave. language	0.701	0.660	0.750	0.697
Doc-BERT	CoMMuTE	0.713	0.738	0.763	0.738
	test2017	-	0.730	0.820	0.775
	test2016	0.759	0.789	0.837	0.795
	MLT	-	0.582	0.716	0.649
	Ave. language	0.745	0.710	0.784	0.739
COMET	CoMMuTE	0.845	0.819	0.818	0.827
	test2017	-	0.810	0.839	0.824
	test2016	0.876	0.851	0.864	0.864
	MLT	-	0.697	0.785	0.741
	Ave. language	0.861	0.794	0.827	0.814
Doc-COMET	CoMMuTE	0.750	0.729	0.724	0.734
	test2017	-	0.744	0.761	0.753
	test2016	0.814	0.776	0.790	0.796
	MLT	-	0.624	0.703	0.663
	Ave. language	0.782	0.718	0.745	0.737

Appendix D

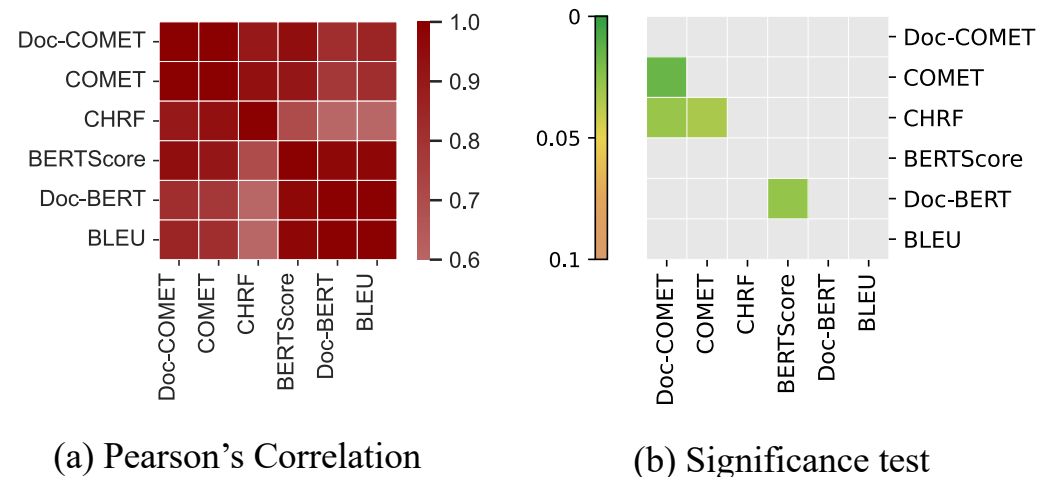


Figure A2. (a) Pearson's correlation between automatic evaluation metrics. (b) p -values from Williams significance tests. Coloured cells indicate that the correlation with human scores of the metric in the row is significantly higher than that of the metric in the column. Results are reported on the THumB dataset.

References

- Ataman, D.; Birch, A.; Habash, N.; Federico, M.; Koehn, P.; Cho, K. Machine Translation in the Era of Large Language Models: A Survey of Historical and Emerging Problems. *Information* **2025**, *16*, 723. [\[CrossRef\]](#)
- Bapna, A.; Cherry, C.; Zhang, Y.; Jia, Y.; Johnson, M.; Cheng, Y.; Khanuja, S.; Riesa, J.; Conneau, A. mslam: Massively multilingual joint pre-training for speech and text. *arXiv* **2022**, arXiv:2202.01374.
- Li, Y.; Shimizu, S.; Gu, W.; Chu, C.; Kurohashi, S. VISA: An Ambiguous Subtitles Dataset for Visual Scene-aware Machine Translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*; European Language Resources Association: Marseille, France, 2022; pp. 6735–6743.
- Feng, Y.; Li, C.; He, J.; Hou, Z.; Ng, V. Multimodal Neural Machine Translation: A Survey of the State of the Art. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Suzhou, China, 2025; pp. 22141–22158. [\[CrossRef\]](#)
- Sulubacak, U.; Caglayan, O.; Grönroos, S.A.; Rouhe, A.; Elliott, D.; Specia, L.; Tiedemann, J. Multimodal machine translation through visuals and speech. *Mach. Transl.* **2020**, *34*, 97–147. [\[CrossRef\]](#)
- Specia, L.; Frank, S.; Sima'an, K.; Elliott, D. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation*; Shared Task Papers: Berlin, Germany, 2016; Volume 2, pp. 543–553. [\[CrossRef\]](#)
- Caglayan, O.; Barrault, L.; Bougares, F. Multimodal attention for neural machine translation. *arXiv* **2016**, arXiv:1609.03976.
- Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; Specia, L. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. *arXiv* **2017**, arXiv:1710.07177. [\[CrossRef\]](#)
- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318. [\[CrossRef\]](#)
- Banerjee, S.; Lavie, A. 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*; Association for Computational Linguistics: Ann Arbor, MI, USA, 2005; pp. 65–72.
- Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 2685–2702. [\[CrossRef\]](#)
- Lala, C. Multimodal Word Sense Translation. Ph.D. Thesis, University of Sheffield, South Yorkshire, UK, 2021.
- Futeral, M.; Schmid, C.; Laptev, I.; Sagot, B.; Bawden, R. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. *arXiv* **2022**, arXiv:2212.10140.

14. 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. 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)*; Association for Computational Linguistics: Philadelphia, PA, USA, 2021; pp. 6153–6166. [\[CrossRef\]](#)
15. Lee, H.; Yoon, S.; Dérnoncourt, F.; Kim, D.S.; Bui, T.; Jung, K. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 34–39. [\[CrossRef\]](#)
16. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv* **2021**, arXiv:2104.08718.
17. Maeda, K.; Kurita, S.; Miyanishi, T.; Okazaki, N. Vision Language Model-based Caption Evaluation Method Leveraging Visual Context Extraction. *arXiv* **2024**, arXiv:2402.17969. [\[CrossRef\]](#)
18. Jiang, Y.E.; Liu, T.; Ma, S.; Zhang, D.; Yang, J.; Huang, H.; Sennrich, R.; Cotterell, R.; Sachan, M.; Zhou, M. BlonDe: An automatic evaluation metric for document-level machine translation. *arXiv* **2021**, arXiv:2103.11878.
19. Vernikos, G.; Thompson, B.; Mathur, P.; Federico, M. Embarrassingly easy document-level mt metrics: How to convert any pretrained metric into a document-level metric. *arXiv* **2022**, arXiv:2209.13654.
20. Raunak, V.; Kocmi, T.; Post, M. SLIDE: Reference-free Evaluation for Machine Translation using a Sliding Document Window. *arXiv* **2024**, arXiv:2309.08832. [\[CrossRef\]](#)
21. Castilho, S.; Popović, M.; Way, A. On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*; European Language Resources Association: Marseille, France, 2020; pp. 3735–3742.
22. Haq, S.; Osuji, C.; Castilho, S.; Davis, B.; Ferreira, T.C. Long-context Reference-based MT Quality Estimation. In *Proceedings of the Tenth Conference on Machine Translation*; Association for Computational Linguistics: Suzhou, China, 2025; pp. 905–912.
23. Furuse, O.; Iida, H. An example-based method for transfer-driven machine translation. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Montreal, QC, Canada, 25–27 June 1992.
24. Chiang, D. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (acl'05)*; Association for Computational Linguistics: Ann Arbor, MI, USA, 2005; pp. 263–270.
25. Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
26. Wu, Y. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
27. Vaswani, A. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)**; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 6000–6010.
28. Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; et al. Findings of the WMT24 general machine translation shared task: The LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 1–46.
29. Castilho, S.; Moorkens, J.; Gaspari, F.; Sennrich, R.; Sosoni, V.; Georgakopoulou, P.; Lohar, P.; Way, A.; Miceli-Barone, A.V.; Gialama, M. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of the Machine Translation Summit XVI: Research Track*; Association for Machine Translation: Nagoya, Japan, 2017; pp. 116–131.
30. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus phrase-based machine translation quality: A case study. *arXiv* **2016**, arXiv:1608.04631.
31. Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; Yu, D. MM-LLMs: Recent Advances in MultiModal Large Language Models. *arXiv* **2024**, arXiv:2401.13601. [\[CrossRef\]](#)
32. Barrault, L.; Bougares, F.; Specia, L.; Lala, C.; Elliott, D.; Frank, S. Findings of the Third Shared Task on Multimodal Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 304–323. [\[CrossRef\]](#)
33. Barrault, L.; Chung, Y.A.; Meglioli, M.C.; Dale, D.; Dong, N.; Duquenne, P.A.; Elsahar, H.; Gong, H.; Heffernan, K.; Hoffman, J.; et al. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv* **2023**, arXiv:2308.11596.
34. Lala, C.; Specia, L. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 7–12 May 2018.
35. Ji, B.; Zhang, T.; Zou, Y.; Hu, B.; Shen, S. Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective. *arXiv* **2022**, arXiv:2210.08478. [\[CrossRef\]](#)
36. Elliott, D.; Frank, S.; Sima'an, K.; Specia, L. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 70–74. [\[CrossRef\]](#)
37. Elliott, D. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 2974–2978.

38. Futeral, M.; Schmid, C.; Sagot, B.; Bawden, R. Towards Zero-Shot Multimodal Machine Translation. *arXiv* **2024**, arXiv:2407.13579. [[CrossRef](#)]
39. Calixto, I.; Stein, D.; Matusov, E.; Lohar, P.; Castilho, S.; Way, A. Using Images to Improve Machine-Translating E-Commerce Product Listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*; Association for Computational Linguistics: Valencia, Spain, 2017; pp. 637–643.
40. Calixto, I.; Liu, Q.; Campbell, N. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv* **2017**, arXiv:1702.01287.
41. Wang, D.; Xiong, D. Efficient Object-Level Visual Context Modeling for Multimodal Machine Translation: Masking Irrelevant Objects Helps Grounding. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2720–2728. [[CrossRef](#)]
42. Lin, H.; Meng, F.; Su, J.; Yin, Y.; Yang, Z.; Ge, Y.; Zhou, J.; Luo, J. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1320–1329.
43. Caglayan, O.; Madhyastha, P.; Specia, L.; Barrault, L. Probing the need for visual context in multimodal machine translation. *arXiv* **2019**, arXiv:1903.08678.
44. Yao, S.; Wan, X. Multimodal Transformer for Multimodal Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 4346–4350. [[CrossRef](#)]
45. Helcl, J.; Libovický, J.; Variš, D. CUNI system for the WMT18 multimodal translation task. *arXiv* **2018**, arXiv:1811.04697.
46. Li, B.; Lv, C.; Zhou, Z.; Zhou, T.; Xiao, T.; Ma, A.; Zhu, J. On vision features in multimodal machine translation. *arXiv* **2022**, arXiv:2203.09173.
47. Caglayan, O. Multimodal Machine Translation. Ph.D. Thesis, Le Mans Université, Le Mans, France, 2019.
48. Elliott, D.; Kádár, A. Imagination improves multimodal translation. *arXiv* **2017**, arXiv:1705.04350.
49. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
50. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; pp. 19730–19742.
51. Gupta, D.; Kharbanda, S.; Zhou, J.; Li, W.; Pfister, H.; Wei, D. CLIPTrans: Transferring visual knowledge with pre-trained models for multimodal machine translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Paris, France, 2023; pp. 2875–2886.
52. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736. [[CrossRef](#)]
53. Gwinnup, J.; Duh, K. A Survey of Vision-Language Pre-training from the Lens of Multimodal Machine Translation. *arXiv* **2023**, arXiv:2306.07198. [[CrossRef](#)]
54. Caglayan, O.; Kuyu, M.; Amac, M.S.; Madhyastha, P.; Erdem, E.; Erdem, A.; Specia, L. 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*; Association for Computational Linguistics: Philadelphia, PA, USA, 2021; pp. 1317–1324. [[CrossRef](#)]
55. Graham, Y.; Baldwin, T.; Moffat, A.; Zobel, J. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*; Association for Computational Linguistics: Sofia, Bulgaria, 2013; pp. 33–41.
56. Bojar, O.; Federmann, C.; Haddow, B.; Koehn, P.; Post, M.; Specia, L. Ten years of WMT evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop “Translation Evaluation—From Fragmented Tools and Data Sets to an Integrated Ecosystem”*; RWTH Aachen University: Portorož, Slovenia, 2016; pp. 27–34.
57. Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. Approaches to human and machine Translation Quality Assessment. In *Translation Quality Assessment: From Principles to Practice*; Machine Translation: Technologies and Applications; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 1, pp. 9–38.
58. Graham, Y.; Baldwin, T.; Moffat, A.; Zobel, J. Can machine translation systems be evaluated by the crowd alone. *Nat. Lang. Eng.* **2017**, *23*, 3–30. [[CrossRef](#)]
59. Kocmi, T.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Gowda, T.; Graham, Y.; Grundkiewicz, R.; Haddow, B.; et al. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 1–45.
60. Lommel, A.; Uszkoreit, H.; Burchardt, A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica* **2014**, *12*, 455–463. [[CrossRef](#)]
61. Kocmi, T.; Zouhar, V.; Avramidis, E.; Grundkiewicz, R.; Karpinska, M.; Popović, M.; Sachan, M.; Shmatova, M. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. *arXiv* **2024**, arXiv:2406.11580.
62. Rossi, C.; Carré, A. How to choose a suitable neural machine translation solution: Evaluation of MT quality. *Mach. Transl. Everyone Empower. Users Age Artif. Intell.* **2022**, *18*, 51–80.

63. O'Brien, S. How to deal with errors in machine translation: Postediting. *Mach. Transl. Everyone Empower. Users Age Artif. Intell.* **2022**, *18*, 105–106.
64. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. 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*; Association for Machine Translation in the Americas: Cambridge, MA, USA, 2006; pp. 223–231.
65. Bojar, O.; Graham, Y.; Kamran, A.; Stanojević, M. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*; Association for Computational Linguistics: Berlin, Germany; pp. 199–231.
66. Abercrombie, G.; Dinkar, T.; Curry, A.C.; Rieser, V.; Hovy, D. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. In *Proceedings of the the 4th Workshop on Perspectivist Approaches to NLP*; Association for Computational Linguistics: Suzhou, China, 2025; pp. 63–74.
67. Kocmi, T.; Artemova, E.; Avramidis, E.; Bawden, R.; Bojar, O.; Dranch, K.; Dvorkovich, A.; Dukanov, S.; Fishel, M.; Freitag, M.; et al. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation, Suzhou, China, 8–9 November 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.
68. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 392–395.
69. Freitag, M.; Rei, R.; Mathur, N.; Lo, C.k.; Stewart, C.; Avramidis, E.; Kocmi, T.; Foster, G.; Lavie, A.; Martins, A.F.T. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 46–68.
70. Callison-Burch, C.; Fordyce, C.S.; Koehn, P.; Monz, C.; Schroeder, J. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 136–158.
71. Freitag, M.; Grangier, D.; Caswell, I. BLEU might be guilty but references are not innocent. *arXiv* **2020**, arXiv:2004.06063.
72. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
73. Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2002; pp. 138–145.
74. Hanna, M.; Bojar, O. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*; Association for Computational Linguistics: Philadelphia, PA, USA, 2021; pp. 507–517.
75. Sun, T.; He, J.; Qiu, X.; Huang, X. BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation. *arXiv* **2022**, arXiv:2210.07626. [[CrossRef](#)]
76. Sellam, T.; Das, D.; Parikh, A. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 7881–7892. [[CrossRef](#)]
77. Rei, R.; C. de Souza, J.G.; Alves, D.; Zerva, C.; Farinha, A.C.; Glushkova, T.; Lavie, A.; Coheur, L.; Martins, A.F.T. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 578–585.
78. Fomicheva, M.; Sun, S.; Yankovskaya, L.; Blain, F.; Guzmán, F.; Fishel, M.; Aletras, N.; Chaudhary, V.; Specia, L. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 539–555. [[CrossRef](#)]
79. Yuan, W.; Neubig, G.; Liu, P. Bartscore: Evaluating generated text as text generation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 27263–27277.
80. Thompson, B.; Post, M. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 90–121. [[CrossRef](#)]
81. Rei, R.; Farinha, A.C.; Zerva, C.; van Stigt, D.; Stewart, C.; Ramos, P.; Glushkova, T.; Martins, A.F.T.; Lavie, A. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*; Association for Computational Linguistics: Philadelphia, PA, USA, 2021; pp. 1030–1040.
82. Specia, L.; Harris, K.; Blain, F.; Burchardt, A.; Macketanz, V.; Skadin, I.; Negri, M.; Turchi, M. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of the Proceedings of Machine Translation Summit XVI: Research Track*; Association for Machine Translation: Nagoya, Japan, 2017; pp. 55–71.
83. Guerreiro, N.M.; Rei, R.; Stigt, D.v.; Coheur, L.; Colombo, P.; Martins, A.F. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 979–995. [[CrossRef](#)]

84. Lavie, A.; Hanneman, G.; Agrawal, S.; Kanojia, D.; kiu Lo, C.; Zouhar, V.; Blain, F.; Zerva, C.; Avramidis, E.; Deoghare, S.; et al. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation, Suzhou, China, 8–9 November 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.
85. Kocmi, T.; Federmann, C. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. *arXiv* **2023**, arXiv:2310.13988.
86. Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; Zhou, J. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *arXiv* **2023**, arXiv:2303.04048.
87. Moghe, N.; Fazla, A.; Amrhein, C.; Kocmi, T.; Steedman, M.; Birch, A.; Sennrich, R.; Guillou, L. Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Comput. Linguist.* **2024**, *51*, 73–137. [[CrossRef](#)]
88. Deutsch, D.; Juraska, J.; Finkelstein, M.; Freitag, M. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. *arXiv* **2023**, arXiv:2308.13506.
89. Castilho, S.; Knowles, R. A survey of context in neural machine translation and its evaluation. *Nat. Lang. Process.* **2024**, *31*, 986–1016. [[CrossRef](#)]
90. Maruf, S.; Saleh, F.; Haffari, G. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [[CrossRef](#)]
91. Läubli, S.; Sennrich, R.; Volk, M. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 4791–4796. [[CrossRef](#)]
92. Läubli, S.; Castilho, S.; Neubig, G.; Sennrich, R.; Shen, Q.; Toral, A. A set of recommendations for assessing human–machine parity in language translation. *J. Artif. Intell. Res.* **2020**, *67*, 653–672. [[CrossRef](#)]
93. Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; et al. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*; Association for Computational Linguistics: Singapore, 2023; pp. 1–42. [[CrossRef](#)]
94. Barrault, L.; Biesialska, M.; Bojar, O.; Costa-jussà, M.R.; Federmann, C.; Graham, Y.; Grundkiewicz, R.; Haddow, B.; Huck, M.; Joanis, E.; et al. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 1–55.
95. Chen, D.; Chen, R.; Zhang, S.; Liu, Y.; Wang, Y.; Zhou, H.; Zhang, Q.; Zhou, P.; Wan, Y.; Sun, L. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. *arXiv* **2024**, arXiv:2402.04788.
96. Fernandes, P.; Deutsch, D.; Finkelstein, M.; Riley, P.; Martins, A.F.T.; Neubig, G.; Garg, A.; Clark, J.H.; Freitag, M.; Firat, O. The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. *arXiv* **2023**, arXiv:2308.07286.
97. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [[CrossRef](#)]
98. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [[CrossRef](#)]
99. Koehn, P.; Knowles, R. Six challenges for neural machine translation. *arXiv* **2017**, arXiv:1706.03872.
100. Raunak, V.; Menezes, A.; Junczys-Dowmunt, M. The curious case of hallucinations in neural machine translation. *arXiv* **2021**, arXiv:2104.06683.
101. Sperber, M.; Bojar, O.; Haddow, B.; Javorský, D.; Ma, X.; Negri, M.; Niehues, J.; Polák, P.; Salesky, E.; Sudoh, K.; et al. Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation. *arXiv* **2024**, arXiv:2406.03881.
102. Li, J.; Ataman, D.; Sennrich, R. Vision Matters When It Should: Sanity Checking Multimodal Machine Translation Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 8556–8562. [[CrossRef](#)]
103. Marie, B.; Fujita, A.; Rubino, R. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 566–577. [[CrossRef](#)]
104. Shen, H.; Shao, L.; Li, W.; Lan, Z.; Liu, Z.; Su, J. A Survey on Multi-modal Machine Translation: Tasks, Methods and Challenges. *arXiv* **2024**, arXiv:2405.12669.
105. Frank, S.; Elliott, D.; Specia, L. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Nat. Lang. Eng.* **2018**, *24*, 393–413. [[CrossRef](#)]
106. Raunak, V.; Choe, S.K.; Lu, Q.; Xu, Y.; Metze, F. On Leveraging the Visual Modality for Neural Machine Translation. In *Proceedings of the 12th International Conference on Natural Language Generation*; Association for Computational Linguistics: Tokyo, Japan, 2019; pp. 147–151. [[CrossRef](#)]
107. Gella, S.; Elliott, D.; Keller, F. Cross-lingual Visual Verb Sense Disambiguation. *arXiv* **2019**, arXiv:1904.05092. [[CrossRef](#)]

108. Wang, J.; Madhyastha, P.; Figueiredo, J.; Lala, C.; Specia, L. MultiSubs: A Large-scale Multimodal and Multilingual Dataset. *arXiv* **2022**, arXiv:2103.01910. [[CrossRef](#)]
109. Haq, S.; Huidrom, R.; Castilho, S. DCU ADAPT at WMT24: English to Low-resource Multi-Modal Translation Task. In *Proceedings of the Ninth Conference on Machine Translation*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 810–814. [[CrossRef](#)]
110. Chen, J.; Guo, H.; Yi, K.; Li, B.; Elhoseiny, M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Washington, DC, USA, 2022; pp. 18009–18019.
111. Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. Gpt-4o system card. *arXiv* **2024**, arXiv:2410.21276.
112. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 8748–8763.
113. Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. Gemma 3 technical report. *arXiv* **2025**, arXiv:2503.19786.
114. Bucciarelli, D.; Moratelli, N.; Cornia, M.; Baraldi, L.; Cucchiara, R. Personalizing multimodal large language models for image captioning: An Experimental analysis. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 351–368.
115. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
116. Kasai, J.; Sakaguchi, K.; Dunagan, L.; Morrison, J.; Le Bras, R.; Choi, Y.; Smith, N.A. Transparent Human Evaluation for Image Captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Seattle, WA, USA, 2022; pp. 3464–3478. [[CrossRef](#)]
117. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
118. Gella, S.; Lapata, M.; Keller, F. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv* **2016**, arXiv:1603.09188.
119. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Washington, DC, USA, 2015; pp. 2641–2649.
120. Bawden, R.; Sennrich, R.; Birch, A.; Haddow, B. Evaluating discourse phenomena in neural machine translation. *arXiv* **2017**, arXiv:1711.00513.
121. Tiedemann, J.; Thottingal, S. OPUS-MT—Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*; European Association for Machine Translation: Lisboa, Portugal, 2020; pp. 479–480.
122. Aryabumi, V.; Dang, J.; Talupuru, D.; Dash, S.; Cairuz, D.; Lin, H.; Venkitesh, B.; Smith, M.; Marchisio, K.; Ruder, S.; et al. Aya 23: Open Weight Releases to Further Multilingual Progress. *arXiv* **2024**, arXiv:2405.15032. [[CrossRef](#)]
123. Freitag, M.; Mathur, N.; Deutsch, D.; Lo, C.K.; Avramidis, E.; Rei, R.; Thompson, B.; Blain, F.; Kocmi, T.; Wang, J.; et al. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 47–81. [[CrossRef](#)]
124. Kocmi, T.; Federmann, C.; Grundkiewicz, R.; Junczys-Dowmunt, M.; Matsushita, H.; Menezes, A. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv* **2021**, arXiv:2107.10821.
125. Zerva, C.; Blain, F.; C. De Souza, J.G.; Kanojia, D.; Deoghare, S.; Guerreiro, N.M.; Attanasio, G.; Rei, R.; Orasan, C.; Negri, M.; et al. Findings of the Quality Estimation Shared Task at WMT 2024: Are LLMs Closing the Gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 82–109. [[CrossRef](#)]
126. Graham, Y.; Haddow, B.; Koehn, P. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 72–81. [[CrossRef](#)]
127. Costa-Jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; et al. No language left behind: Scaling human-centered machine translation. *arXiv* **2022**, arXiv:2207.04672.
128. Freitag, M.; Rei, R.; Mathur, N.; Lo, C.k.; Stewart, C.; Foster, G.; Lavie, A.; Bojar, O. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*; Association for Computational Linguistics: Philadelphia, PA, USA, 2021; pp. 733–774.
129. Mathur, N.; Baldwin, T.; Cohn, T. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Philadelphia, PA, USA, 2020; pp. 4984–4997. [[CrossRef](#)]

130. Graham, Y.; Baldwin, T. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 172–176.
131. Moghe, N.; Sherborne, T.; Steedman, M.; Birch, A. Extrinsic Evaluation of Machine Translation Metrics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 13060–13078. [[CrossRef](#)]
132. Chauhan, S.; Daniel, P. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Process. Lett.* **2023**, *55*, 12663–12717.
133. Kaster, M.; Zhao, W.; Eger, S. Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 8912–8925. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.