Assessing the Role of Imagery in Multimodal Machine Translation

Nicholas Kashani Motlagh[1], Jim Davis[1], Tim Anderson[2], Jeremy Gwinnup[2], Grant Erdmann[2]
[1]Department of Computer Science and Engineering, Ohio State University
{kashanimotlagh.1, davis.1719}@osu.edu
[2]Air Force Research Laboratory
{timothy.anderson.20, jeremy.gwinnup.1, grant.erdmann}@us.af.mil

Abstract

In Multimodal Machine Translation (MMT), the use of visual data has shown only marginal improvements compared to text-only models. Previously, the CoMMuTE dataset and associated metric were proposed to score models on tasks where the imagery is necessary to disambiguate between two possible translations for each ambiguous source sentence. In this work, we introduce new metrics within the CoMMuTE domain to provide deeper insights into image-aware translation models. Our proposed metrics differ from the previous CoMMuTE scoring method by 1) assessing the impact of multiple images on individual translations and 2) evaluating a model's ability to jointly select each translation for each image context. Our results challenge the conventional views of poor visual comprehension capabilities of MMT models and show that models can indeed meaningfully interpret visual information, though they may not leverage it sufficiently in the final decision.

1 Introduction

The use of multimodal data, combining visual and textual inputs, is becoming increasingly important in deep learning, especially in language modeling. Multimodal Machine Translation (MMT) presents a unique challenge in this area, as previous Machine Translation (MT) systems traditionally relied only on text. Despite the potential benefits of incorporating imagery, its efficacy in MMT remains controversial. Critics often view imagery as merely a regularizer rather than a core component of translation systems (Caglayan et al., 2016; Wu et al., 2021). This skepticism is fueled by results with the assumption that textual context alone suffices for most translation tasks (Caglayan et al., 2019).

To explore these concerns, the CoMMuTE dataset was developed to test MMT models on source sentences where visual context is essential for accurate selection between possible translations (Futeral et al., 2023). Their proposed evaluation metric scores a model's *preference/choice between two reference translations*, diverging from traditional metrics such as BLEU (Papineni et al., 2002) and Meteor (Banerjee and Lavie, 2005) that instead compare a generated translation against a single reference. Initial analyses using the CoMMuTE dataset and metric indicate that current models show only slight, or no, improvement over using text-only models (Futeral et al., 2023).

Building on this recent foundation, we introduce a new complementary evaluative CoMMuTE metric that assesses a model's understanding of varying imagery on *a fixed reference translation* (as described above in (Futeral et al., 2023)). We additionally provide two group metrics designed to evaluate a model's ability to jointly choose each translation given their associated image contexts.

Results with our proposed metrics demonstrate that in many circumstances, models can indeed effectively understand and properly interpret the visual information, even if the final translation decisions are unaffected. This suggests the significant potential for improvements in model design to further leverage visual information.

2 Related Work

In this section, we present an overview of recent advancements and methodologies in two critical areas of related research. We first explore how imagery can enhance translation capabilities in MMT and subsequently shift our focus to contrastive evaluation methods, which represent a shift from traditional single-reference comparisons to more nuanced assessments using multiple contrasting references.

2.1 Multimodal Machine Translation

MMT typically trains with datasets such as Multi30k (Elliott et al., 2016) to enhance trans-

Proceedings of the Ninth Conference on Machine Translation, pages 1428–1439
November 15-16, 2024 ©2024 Association for Computational Linguistics

lation capabilities, yet results are not largely improved with sufficient textual context (Caglayan et al., 2019). Research such as Elliott (2018) demonstrates that the replacement of associated images with random counterparts often does not significantly impact translation quality, suggesting a predominant reliance on textual data. A later study further indicated that imagery typically serves merely as a form of regularization in training current models (Wu et al., 2021).

When imagery is available at inference time, approaches such as Graph-MMT (Yin et al., 2020), VTLM (Caglayan et al., 2021), Gated Fusion (Wu et al., 2021), and VGAMT (Futeral et al., 2023) are applicable. These methods leverage diverse global visual features from sources such as ResNet-50 (He et al., 2016) and CLIP (Radford et al., 2021), as well as visual semantic features through advanced object detectors like MDETR (Kamath et al., 2021).

In scenarios lacking visual data at inference time, innovative models such as CLIP-Trans (Gupta et al., 2023), UVR-NMT (Zhang et al., 2020), and ImagiT (Long et al., 2021) instead strategically leverage image-text datasets only during their training phase. These models employ sophisticated mechanisms to enhance their semantic understanding during training such as aligning image-text embedding spaces and synthesizing visual features. By pretraining on multimodal data, these models acquire a nuanced understanding of complex semantic relationships that text alone might not fully encapsulate. Some models, such as CLIP-Trans, can be modified to support the use of imagery at inference time by replacing CLIP text embeddings with CLIP image embeddings.

There has also been notable progress in adapting pretrained language models (LMs) such as BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) for multimodal use. Techniques such as visually-conditioned masked language modeling (VMLM) are explored in various architectures (Chen et al., 2020; Lu et al., 2019; Su et al., 2020; Li et al., 2020; Zhou et al., 2021; Ni et al., 2021; Futeral et al., 2023). Furthermore, the development of adapters and other lightweight modules can significantly enhance multimodal capabilities of LMs (Houlsby et al., 2019; Eichenberg et al., 2022; Yang et al., 2022; Tsimpoukelli et al., 2021; Sung et al., 2022; Futeral et al., 2023).

2.2 Contrastive Evaluation

Contrastive evaluation methodologies have become crucial for nuanced assessments of translation systems. These methodologies utilize contrastive test sets designed to challenge models to correctly rank pairs of translations, helping distinguish between correct and incorrect alternatives (Futeral et al., 2023). Contrastive datasets have been used to evaluate linguistic phenomena including grammaticality (Sennrich, 2017), pronoun translation (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019), and multi-sense word disambiguation (Rios Gonzales et al., 2017; Raganato et al., 2019; Futeral et al., 2023). Moreover, the coherence of lexical usage across translations has been thoroughly explored (Bawden et al., 2018; Voita et al., 2019).

3 CoMMuTE Dataset and Metric

The CoMMuTE dataset (Futeral et al., 2023) was recently introduced to score an MMT model's preference between two given translations for an ambiguous source based on the provided imagery. Specifically, CoMMuTE is comprised of 154 ambiguous English sentences, each paired with two contrasting images and their respective translations, where the two translations are available in French, German, and Czech. Each instance in the dataset is structured as a tuple (s, i^a, t^a, i^b, t^b), where s is an ambiguous source sentence and (i^a, i^b) are images that disambiguate the sentence into two possible translations (t^a, t^b), respectively. For example, in Fig. 1, the English source sentence "That's lots of bucks!" could refer to either deer or dollars, and the image is needed to determine the appropriate context.

To specifically score such disambiguation capabilities, the authors proposed a metric, which we refer to as TextCoMMuTE (TC), that compares the model's preference for the correct translation over the incorrect translation based on a single provided image context.

The model's uncertainty in a translation t given a source s and an image i is quantified by perplexity, defined as

$$\mathcal{P}(s, i, t) = \exp\left(-\frac{1}{N}\sum_{k=1}^{N} \log p(t_k|s, i, t_{<k})\right) \quad (1)$$

Here, N is the number of tokens in the translation, t_k is the k-th token in the translation, and $p(t_k|s, i, t_{<k})$ denotes the conditional probability

of the k-th token given the source, image, and preceding tokens. In practice, this probability is approximated using the softmax of model outputs. Perplexity can be seen as a measure of uncertainty as it is the exponential of the negative mean log probability. Hence, *lower* perplexity is desired for a correct output versus an incorrect output.

The TC metric is then defined for a single image-translation triple (i^m, t^m, t^n) as

$$TC^{m,n} = \mathbb{1}\{\mathcal{P}(s, i^m, t^m) < \mathcal{P}(s, i^m, t^n)\} \quad (2)$$

where i^m and t^m correspond to the matching image/translation and t^n is the incorrect translation in the associated triple. Moreover, $\mathbb{1}$ is the indicator function that is 1 if the perplexity for the correct translation is less than that of the incorrect translation, and 0 otherwise.

Note that each of the 154 tuples in CoMMuTE yields 2 TC scores: $TC^{a,b}$ and $TC^{b,a}$. Hence, there are actually 308 individual TC scores for the dataset. An average is taken over the N=154 TC pairs as a summary statistic

$$TC = \frac{1}{2N} \sum_{j=1}^{N} \{TC^{a_j,b_j} + TC^{b_j,a_j}\} \quad (3)$$

Again, the TC score (Eqn. 3) views the two triples in each tuple *independently* even though both triples are associated with the same source sentence. TC scores range from 0-1 with 1 indicating correct disambiguation of all triples in the dataset. A text-only model scores a TC of 0.5 by definition (assuming no ties in perplexity) because for any tuple j in the dataset, exactly one of TC^{a_j,b_j} and TC^{b_j,a_j} will be 1 while the other is 0 (*i.e.*, the image makes no contribution to the translation preference for a given source).

From an MMT perspective, this metric is insightful as translations with lower perplexities are typically more likely to be generated or appear higher in an n-best list.

4 Enhanced CoMMuTE Metrics

We now propose new complementary contrastive metrics to provide a more nuanced understanding of the interpretation of imagery for models with the CoMMuTE dataset.

4.1 ImageCoMMuTE

Rather than comparing two translations with the same image and source as is done with TC, we in-

(a) French Translation a: *Il y a beaucoup de cerfs !*

(b) French Translation b: *Cela fait beaucoup de dollars !*

Figure 1: English Source: *That's lots of bucks!*

stead examine the contribution of two *different* images to the *same* translation. From this perspective, we can directly assess whether the correctly associated image is appropriately affecting model uncertainty (reducing the perplexity of its corresponding translation). For a source s, images (i^m, i^n), and a translation t^m, we define ImageCoMMuTE (IC) as

$$IC^{m,n} = \mathbb{1}\{\mathcal{P}(s, i^m, t^m) < \mathcal{P}(s, i^n, t^m)\} \quad (4)$$

where i^m is the correctly associated image and i^n is incorrectly associated image for translation t^m. Similar to TC, one can aggregate scores over a dataset by taking the mean of the N=154 pairs

$$IC = \frac{1}{2N} \sum_{j=1}^{N} \{IC^{a_j,b_j} + IC^{b_j,a_j}\} \quad (5)$$

Scores for IC range from 0-1, and a score of 0.5 indicates a random preference for the image context.

Our IC metric evaluates changes in model confidence for the *same* translation when presented with varying imagery. This approach directly assesses the interplay between imagery and text interpretation within the model. This differs from the work presented in Elliott (2018), where they assess average differences in model uncertainty, while we assess indicators of decisions. This IC metric also alleviates any possible concerns of the reliance on comparing perplexity averages and calibration *across* translations (as is done with TC). We will return to these potential issues in our discussion later. By maintaining a single reference translation across different visual contexts, our IC metric provides a more robust and precise measure of how imagery is understood by the model.

4.2 Group CoMMuTE

Though TC and IC are insightful metrics on their own, they both ignore the consistency desired for the underlying source-translation *pairs*. With TC, the set of both *translations* is independently processed twice (each time with a different image context). Similarly with IC, the set of both *images* is independently processed twice (each time with a different translation target). What is truly desired is that the model consistently and correctly understands *both* cases for each set jointly to demonstrate true understanding.

Therefore, we propose a new group variant for TC and IC. To evaluate consistency across the paired nature of the task, we define Group TextCoMMuTE (GTC) as

$$GTC^{a,b} = TC^{a,b} \cdot TC^{b,a} \tag{6}$$

and Group ImageCoMMuTE (GIC) as

$$GIC^{a,b} = IC^{a,b} \cdot IC^{b,a} \tag{7}$$

These group metrics function with a logical "AND" between the two independent triple scores, ensuring that a score of 1 reflects consistent and correct interpretations for the tuple as a whole. As earlier, one can also aggregate group scores using a mean with

$$GTC = \frac{1}{N}\sum_{j=1}^{N} GTC^{a_j,b_j} \tag{8}$$

$$GIC = \frac{1}{N}\sum_{j=1}^{N} GIC^{a_j,b_j} \tag{9}$$

These scores also yield values between 0-1.

Our primary goal is to assess if the model properly interprets and understands imagery for the translations. Group scores such as GTC and GIC are crucial because they assess consistent model behavior with different text-image combinations, indicating true comprehension rather than coincidental correctness.

5 Experiments and Results

We present a comprehensive assessment of the previous and new CoMMuTE metrics on three pretrained English-to-French MMT models. Our evaluation is structured to elucidate how well these models understand the imagery with respect to resolving ambiguities in the CoMMuTE dataset. We begin by evaluating performance on the original CoMMuTE dataset, followed by an assessment using an extended set of imagery we collected for each CoMMuTE tuple to reveal further strengths and weaknesses across models.

5.1 Models

We employed three English-to-French MMT models, each chosen for its unique approach to integrating visual data with textual information. Across all models, we preprocessed imagery by resizing the smaller edge to 224px (maintaining the aspect ratio) and then taking a center crop of 224px $\times$ 224px.

VGAMT. The authors of CoMMuTE proposed VGAMT (Futeral et al., 2023), enhancing a pretrained mBART MT model (Liu et al., 2020) by incorporating CLIP ViT-B/32 image embeddings and fine-tuning adapters. While VGAMT included an object detector and a visually guided attention mechanism, our evaluation focused on its simplified variant from their ablation study (Futeral et al., 2023), which solely uses CLIP image embeddings. This model was trained using both visual masked language modeling and MMT objectives, having 1B total parameters. In our experiments, we employed three VGAMT models provided by the authors, each trained with a different random seed.

CLIP-Trans. The authors (Gupta et al., 2023) align the embedding spaces of a pretrained mBART MT model (Liu et al., 2020) with a multilingual M-CLIP model (Carlsson et al., 2022) via a mapping network. The model first trains on an image-captioning task using M-CLIP image embeddings followed by text-only MT training with M-CLIP text embeddings. They also suggest that imagery can be utilized at inference time, substituting M-CLIP text embeddings with image embeddings, even though it is not directly trained on MMT. We used a model following this approach with 1.3B total parameters. In the experiments, we evaluated one CLIP-Trans model provided by the authors.

Gated Fusion. This model introduces a dynamic gating mechanism that adaptively combines image and text representations, with gate values ranging from 0 to 1 for image components (Wu et al., 2021). The model leverages ResNet-50 (He et al., 2016) image features and a tiny transformer for a total of 32M parameters (substantially smaller than CLIP-Trans and VGAMT). We trained the model solely on the Multi30K dataset (Elliott et al., 2016), adhering to the authors' training protocol. We observed that the gating mechanism frequently assigns low values, often near 0, which tends to minimize the impact of visual data. To better incorporate image content into the translation process, we trained additional variants with fixed gate values of 0.25, 0.5,

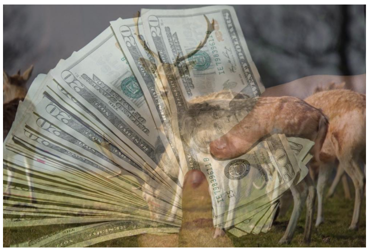

Figure 2: Mixed imagery from Fig. 1 used for a pseudo-text-only baseline.

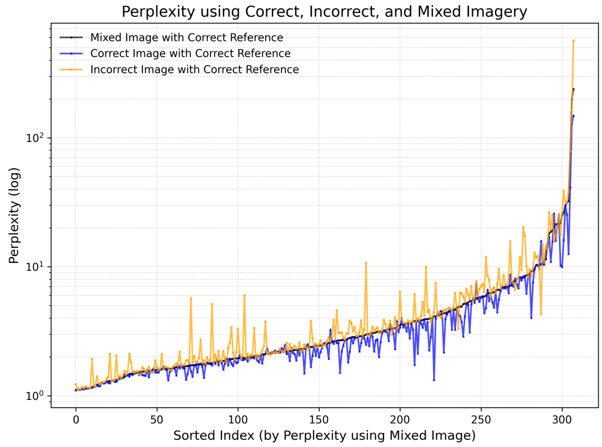

Figure 3: Perplexities of the correct translations using the correct image, the incorrect image, and the mixed image.

and 0.75. Each of these variants was trained and evaluated using three different random seeds.

5.2 Baseline Results

We first conducted a baseline evaluation on the CoMMuTE dataset. The second and third columns in Table 1 display the mean TC and GTC scores taken across models with random seeds (standard deviations were very low in all cases). For reference, a pure text-only MT model will have TC=0.5 and GTC=0, since the model will always choose one translation over the other for each tuple.

VGAMT scores highest in these two metrics, with the CLIP-Trans and Gated Fusion variants scoring near text-only in TC. This model also scores the highest in BLEU on Multi30k, as reported in previous work (Futeral et al., 2023; Gupta et al., 2023). The GTC scores of all models are above 0%, suggesting that all models can consistently disambiguate at least some tuples, though the scores are low. The gate values within the default Gated Fusion model were inspected and found to be near 0 (as expected). Interestingly, we see that TC for Gated Fusion improves slightly with a fixed larger gate value of 0.25 indicating that the strength of imagery does have the potential to change translations.

5.3 Comparison with Ambiguous Imagery

We next examined how much the imagery affected model decisions in comparison to the underlying textual bias. We compared the changes in TC scores using the original image context pairs (from CoMMuTE) versus an ambiguous mixed image.

As MMT models are trained with both imagery and text, one cannot properly obtain a pure text-only result through simple methods such as passing a zero image or removing the image context from the tokens. To obtain a pseudo-text-only baseline, we employed a 50/50% "mixup" (Zhang et al., 2018) of the two image contexts for each tuple to create a single ambiguous image (see Fig. 2). Here, both image contexts are provided in a single image. However, there are other possible ways to create ambiguous imagery, such as arranging the images side-by-side. In Fig. 3, we see the perplexities of the correct translations using the mixed imagery typically fall between the perplexities using the correct and incorrect imagery, supporting the use of the mixed imagery as a baseline for comparison. We evaluated TC using this mixed image and also using the original images to get two competing TC scores for each image-translation triple. Note that the pseudo-text-only MMT model will score TC=0.5 (and GTC=0) by definition (we are using the same mixed image across two comparisons, and thus, preference does not change).

We measure changes in the score between the original images and the mixed image for each tuple using four consistency rates. The first two rates measure the percent of image-translation triples for which the original imagery and the mixed imagery gave different preferences for translations. That is, in these cases, the model's decision when using the original imagery was different from the model's decision when using the mixed imagery. The inconsistent positive rate (IPR) measures the percentage of image-translation triples that chose the right translation with the original imagery and the opposite/wrong translation with mixed imagery. The inconsistent negative rate (INR) measures the percentage of image-translation triples that chose the wrong translation with the original imagery and the opposite/right translation with mixed imagery. The performance of the remaining examples can be

Model	Mean TC ↑	Mean GTC ↑	IPR ↑	INR ↓	CPR ↑	CNR ↓
VGAMT	0.63	0.26	0.13	0.00	0.50	0.37
CLIP-Trans	0.51	0.03	0.01	0.00	0.50	0.49
Gated Fusion	0.50	0.02	0.01	0.01	0.49	0.49
Gated Fusion$_{0.25}$	0.52	0.10	0.07	0.05	0.45	0.43
Gated Fusion$_{0.5}$	0.50	0.07	0.05	0.05	0.45	0.45
Gated Fusion$_{0.75}$	0.49	0.02	0.02	0.04	0.46	0.48

Table 1: Baseline TC and GTC scores on the original CoMMuTE dataset, and consistency rates compared to pseudo-text-only baseline.

quantified by a consistent positive rate (CPR) and a consistent negative rate (CNR), measuring the percentage of triples whose correct and incorrect preferences did not change when using the original or mixed imagery. Since the corpus is evenly split into 2 ambiguities, these rates are bounded in [0, 0.5] with IPR + CNR = INR + CPR = 0.5.

The last four columns in Table 1 display the consistency rates using the pseudo-text-only baseline for each of the models. The VGAMT model scores the highest IPR of 0.13 with an INR of 0, indicating that the model corrected 13% of translations without any negative impact when using the original imagery. In contrast, the CLIP-Trans and Gated Fusion variants show smaller IPR and INR rates, suggesting that imagery has a weaker yet still noticeable effect on these models. The higher INR rates for Gated Fusion models indicate that imagery can actually hurt their performance.

By examining the CPR and CNR rates in the table, we see that imagery may not be significantly impactful in the decisions across all models. These rates only measure the proportion of image-translation triples (with the original imagery) that *agree* with the pseudo-text-only baseline (with the mixed imagery). They do not describe if the model associates correct/incorrect imagery with translation confidence. The model still might correctly associate the original imagery, giving lower perplexity of the correct translation (desired), but this change may not be drastic enough to overturn the model's underlying textual preference. This highlights the need for a metric, such as the proposed IC, to measure how confidence in a translation changes with correct and incorrect imagery.

5.4 ImageCoMMuTE Results

We next conducted an evaluation of the CoMMuTE dataset using our proposed IC and GIC metrics. Table 2 displays the mean IC and GIC scores taken across the models with random seeds. Note that IC and GIC metrics are undefined for a pure text-only MT model, and thus, we cannot compute the four consistency rates.

Model	Mean IC ↑	Mean GIC ↑
VGAMT	0.81	0.66
CLIP-Trans	0.58	0.22
Gated Fusion	0.51	0.11
Gated Fusion$_{0.25}$	0.51	0.12
Gated Fusion$_{0.5}$	0.50	0.13
Gated Fusion$_{0.75}$	0.50	0.11

Table 2: Baseline IC and GIC scores.

Model	TC	IC
VGAMT vs CLIP-Trans	0.39	0.18
VGAMT vs Gated Fusion$_{0.25}$	0.25	0.16
Gated Fusion$_{0.25}$ vs CLIP-Trans	0.36	0.32

Table 3: Intersection-Over-Union of failures as determined by TC and IC.

Our image-based metrics (IC and GIC) demonstrate that VGAMT interprets imagery most effectively, achieving 0.81 on IC and 0.66 on GIC, which are significantly higher than the TC of 0.63 and GTC of 0.26. Other models continue to score only slightly above 0.5. We find that of the models we tested, those that scored highest on MMT quality metrics also scored highest in our proposed metrics (as reported in (Futeral et al., 2023; Gupta et al., 2023)). These results demonstrate that VGAMT more appropriately adjusts uncertainty in a translation based on imagery.

We also investigated whether the different models made the same errors. We identified the image-translation triples where each model made errors in terms of TC and also for IC. We then calculated the intersection-over-union (IOU) between 2 models, which is a set similarity metric defined as the ratio of the number of image-translation triples common to both error sets for a given metric (intersection) to the total number of unique image-translation triples in both error sets (union). This metric helps quantify the similarity in errors across models as a scalar bounded in [0,1] where 1 signifies exact similarity in errors. The results in Table 3 reveal that models do not strongly make the same mistakes yet do share some overlap.

5.5 Extended CoMMuTE

We next extended the CoMMuTE dataset by incorporating additional images per translation in each tuple. This extension allows for a broader assessment of model performance across diverse image inputs and enables a search for images that could either improve or degrade the scores.

For each ambiguous source s, we manually generated two distinct, unambiguous captions, c^a and c^b, which correspond directly to the translations t^a and t^b, respectively. For example, the English sentence "That's lots of bucks!" is transformed to "a photo of deer" and "a photo of dollars".

Utilizing these unambiguous captions, we then sourced corresponding images from the DataComp-12.8M dataset (Gadre et al., 2023), which comprises 12.8 million image-text pairs harvested from the Common Crawl (Common Crawl). The DataComp dataset serves as a foundation dataset for enhancing the training of CLIP models. We employ a CLIP ViT-B/32 model, pretrained on the LAION-5B dataset (Schuhmann et al., 2022), to retrieve images most similar (cosine similarity) to our unambiguous captions.

From this candidate set of imagery, the top 15 images that most closely aligned with each caption, adhering to a minimum dimension of 64 pixels and a maximum aspect ratio of 2.5, were retrieved automatically. We manually selected the four most representative images from this set (due to potentially noisy images retrieved). If fewer than 4 suitable images were found, additional images were sourced from Google Images. This method resulted in a total of 1540 images, providing 5 images (instead of just 1) for each unambiguous translation. Consequently, this extended CoMMuTE dataset includes the original source s, translations t^a and t^b, and now 5 images each for i^a and i^b.

With this extended CoMMuTE dataset, we examined if there existed subsets of imagery that could significantly increase or decrease the GIC score (as we deem GIC the most important metric for each model). For each tuple in our extended dataset, we identified the image pair (one image taken from each image set) that maximizes or minimizes the GIC score. As multiple pairs can meet the criteria, we select the pair that optimizes

$$\begin{aligned}&\{\mathcal{P}(s,i^a,t^a)-\mathcal{P}(s,i^b,t^a)\} \quad + \\ &\{\mathcal{P}(s,i^b,t^b)-\mathcal{P}(s,i^a,t^b)\}\end{aligned} \tag{10}$$

This expression reflects the confidence gaps for the translations. Given that a lower perplexity indicates a better result and considering the ordering of differences in Eqn. 10, we minimize (or maximize) this equation to maximize (or minimize) the GIC score accordingly. When seeking images to maximize the GIC score, we break ties by finding the image pair that *minimizes* Eqn. 10 (can be negative). When seeking images to minimize the GIC score, we break ties with the image pair that *maximizes* Eqn. 10. We refer to the image subset specifically tailored to maximize GIC as Image-Oracle. We also tracked the replacement rate (RR) of the number of images replaced from the original dataset.

As shown in Table 4, the maximal GIC image subsets show high effectiveness, with VGAMT scoring a Max IC of 0.96 and a Max GIC of 0.92. This suggests that the model can accurately interpret the intended visual signals in these particular image pairs for nearly all translations. This is further supported by the notably higher Max IC and GIC scores in the CLIP-Trans and Gated Fusion variants. Conversely, we see that sets of images can be found to hurt performance, especially in CLIP-Trans and Gated Fusion. Examples of replaced imagery can be seen in Fig. 4. Therefore, it is possible to have imagery that drastically improves or degrades the scores. We see that replacement rates are high, indicating that the original dataset is not prominent in these maximal/minimal subsets. The results with maximal/minimal GIC show that the model does indeed have an internal understanding of the imagery with respect to the translation task.

We would expect the Image-Oracle images that maximized GIC to similarly improve TC and GTC scores. However, Table 5 shows only minor improvements in TC and GTC across models. Thus, even though the IC and GIC metrics strongly indicate the image interpretability of the models, the TC and GTC metrics fail to highlight the potential contribution of imagery.

6 Discussion

This study introduced image-based and group metrics for CoMMuTE to better evaluate if models do

Model	Min IC $\uparrow$	Min GIC $\uparrow$	RR	Max IC $\uparrow$	Max GIC $\uparrow$	RR
VGAMT	0.59	0.33	0.80	0.96	0.92	0.71
CLIP-Trans	0.46	0.01	0.77	0.89	0.77	0.77
Gated Fusion	0.40	0.00	0.77	0.73	0.48	0.78
Gated Fusion$_{0.25}$	0.38	0.00	0.80	0.86	0.71	0.80
Gated Fusion$_{0.5}$	0.35	0.00	0.81	0.88	0.76	0.77
Gated Fusion$_{0.75}$	0.37	0.00	0.79	0.85	0.71	0.80

Table 4: Minimum and maximum IC and GIC scores along with replacement rates.

Original

Oracle Best

Oracle Worst

Figure 4: Examples from the CoMMuTE dataset with original imagery (top row), oracle best replacements (middle row), and oracle worst replacements (bottom row) as determined by VGAMT.

Model	Mean TC $\uparrow$	Mean GTC $\uparrow$
VGAMT	0.67	0.34
CLIP-Trans	0.52	0.05
Gated Fusion	0.51	0.02
Gated Fusion$_{0.25}$	0.64	0.28
Gated Fusion$_{0.5}$	0.59	0.18
Gated Fusion$_{0.75}$	0.56	0.12

Table 5: Image-Oracle TC and GTC scores.

Model	Mean TC $\uparrow$	Mean GTC $\uparrow$
VGAMT	0.66	0.32
CLIP-Trans	0.52	0.03
Gated Fusion	0.51	0.01
Gated Fusion$_{0.25}$	0.60	0.21
Gated Fusion$_{0.5}$	0.58	0.15
Gated Fusion$_{0.75}$	0.53	0.07

Table 6: Image-Oracle TC scores with the shared prefix removed in perplexity computation.

understand imagery in MMT. In this section, we explore possible reasons why TC scores are so much lower than IC and discuss future directions on how to further leverage the imagery to improve MMT.

There are two potential issues related to perplexity and calibration that may affect the TC/GTC scores. First, there is an assumption that perplexity is indeed an appropriate uncertainty metric to compare *two* translations. Perplexity is a transform of the mean log probability and, therefore, relies on averages where all tokens are weighted equally (Ueda et al., 2024). There may indeed be other better measures of uncertainty (Kauf and Ivanova, 2023). It is also assumed that the model is well calibrated to properly compare *across* translations.

One method to examine the effects of averages across sequences of different lengths in the perplex-

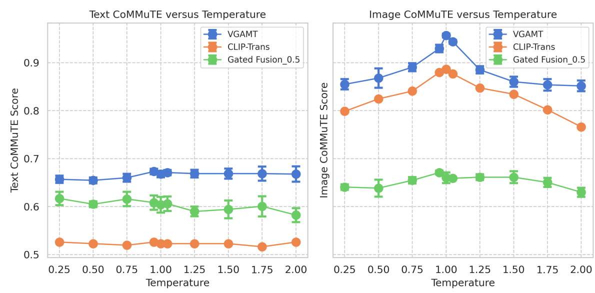

Figure 5: Calibration results using temperature scaling.

ity computation is to remove any shared prefix in t^a, t^b before computing perplexity and then compare to the results without prefix removal (original method). Ignoring common prefixes (while still weighting the remaining tokens equally) actually shows a slight degradation in scores (as illustrated in Table 6). These results suggest perplexity (a transform of mean log probability) does have some issues as a comparison method. However, this does not fully explain the low TC/GTC scores.

We also investigated the effects of model calibration using a simple global temperature scaling method (Guo et al., 2017) across a range of temperature values from 0.25 to 2. As shown in Fig. 5, the TC scores appear unaffected, indicating potential miscalibration, while IC scores suggest that models are relatively well-calibrated (at T=1). We also examined higher temperatures, which did not change the results, suggesting calibration does not appear to be primarily responsible for the TC/GTC degradation.

Therefore, given the stronger results from IC/GIC, we believe the main overall issue with TC/GTC is that the underlying textual preference/bias in these models is too strong and does not allow much influence from the imagery (which we have shown to be interpreted well by the models).

7 Recommendations for Future Work

One future area of work is the integration of imagery *earlier* in the model's architecture rather than appending them at the end of the processing chain (Wu et al., 2021; Gupta et al., 2023). Integrating image features earlier in the model's architecture could enhance the model's ability to better leverage the rich contextual cues provided by the imagery. This approach may result in translations that are more contextually nuanced, with increased attention to specific words critical for disambiguation.

Additionally, enhancing the impact of visual signals *within* the model could also prove beneficial. This could be achieved by adjusting the gate values in models that use gating mechanisms, such as Gated Fusion (Wu et al., 2021), to strengthen the influence of visual data. As demonstrated, setting a fixed gate value that prioritizes visual information could help in situations where visual context is crucial for disambiguating textual content. Even though the non-gated VGAMT was the top performer, there is still room for improvement by strengthening the role of imagery in the processing using some method of gating or amplification.

Earlier we have shown that the IOU of errors between model pairs did not have strong alignment. This diversity implies that ensembling different models could potentially mitigate individual weaknesses and enhance overall performance.

8 Conclusion

Our study challenges the widespread belief that visual cues are not generally very helpful to MMT. By employing our proposed IC and Group CoMMuTE metrics within an expanded CoMMuTE dataset, we have established a robust framework for assessing if visual information is understood in MMT systems. Our results reveal that while visual data does indeed support translation preferences, it is not leveraged significantly to enhance the outcomes over the underlying textual bias. Our findings mark a promising direction for future research in MMT, suggesting that further exploration could uncover ways to amplify this positive impact.

Acknowledgements. We thank the authors of CoMMuTE for providing their CLIP-only VGAMT models. We also thank Logan Frank and Robert Sunderhaft for their assistance.

Limitations

Firstly, we evaluated English-French translations in CoMMuTE. It remains to be seen whether the results generalize to other languages. Additionally, our evaluations were conducted on an extended set of 5 images, whereas larger sets (e.g., 100 images) would provide more robust insights. Furthermore, we relied on the default single reference translation for each image. Having additional translations for each image context would enable a more comprehensive evaluation.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does Multimodality Help Human and Machine for Translation and Image Captioning? In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual Visual Pretraining for Multimodal Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and Multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12375, pages 104–120. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Common Crawl. Common crawl. `https://commoncrawl.org`.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2416–2428, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial Evaluation of Multimodal Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

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

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In Search of the Next Generation of Multimodal Datasets. In *Advances in Neural Information Processing Systems*, volume 36, pages 27092–27112. Curran Associates, Inc.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR. ISSN: 2640-3498.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. CLIPTrans: Transferring Visual Knowledge with Pre-trained Models for Multimodal Machine Translation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2863–2874, Paris, France. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA. IEEE.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR. ISSN: 2640-3498.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, Montreal, QC, Canada. IEEE.

Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12375, pages 121–137. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. Place: Cambridge, MA Publisher: MIT Press.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative Imagination Elevates Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, Online. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3977–3986.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An Open Large-scale Dataset for Training Next Generation Image-Text Models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5227, New Orleans, LA, USA. IEEE.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. 2024. Token-length Bias in Minimal-pair Paradigm Datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16224–16236, Torino, Italia. ELRA and ICCL.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models. In *Advances in Neural Information Processing Systems*.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural Machine Translation with Universal Visual Representation. In *International Conference on Learning Representations*.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4153–4163.