HAL open science

Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, Rachel Bawden

To cite this version:

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, Rachel Bawden. Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul 2023, Toronto, Canada. pp.5394-5413, 10.18653/v1/2023.acl-long.295 . hal-03977982

HAL Id: hal-03977982

https://inria.hal.science/hal-03977982v1

Submitted on 22 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation

Matthieu Futeral[1,2] **Cordelia Schmid**[1,2] **Ivan Laptev**[1,2]
Benoît Sagot[1] **Rachel Bawden**[1]
[1]Inria Paris
[2]Département d'informatique de l'ENS, CNRS, PSL Research University
firstname.lastname@inria.fr

Abstract

One of the major challenges of machine translation (MT) is ambiguity, which can in some cases be resolved by accompanying context such as images. However, recent work in multimodal MT (MMT) has shown that obtaining improvements from images is challenging, limited not only by the difficulty of building effective cross-modal representations, but also by the lack of specific evaluation and training data. We present a new MMT approach based on a strong text-only MT model, which uses neural adapters, a novel guided self-attention mechanism and which is jointly trained on both visually-conditioned masking and MMT. We also introduce CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation set of ambiguous sentences and their possible translations, accompanied by disambiguating images corresponding to each translation. Our approach obtains competitive results compared to strong text-only models on standard English→French, English→German and English→Czech benchmarks and outperforms baselines and state-of-the-art MMT systems by a large margin on our contrastive test set. Our code[1] and CoMMuTE[2] are freely available.

1 Introduction

Multimodal machine translation (MMT) typically refers to the use of additional non-textual data in text-based machine translation (MT). Here, we focus on the case where source texts are accompanied by images, the idea being to exploit visual data to improve the translation of ambiguous sentences. For example, in Figure 1, the English word *glasses* can either be translated as French *verres* 'drinking vessels' or *lunettes* 'spectacles', an ambiguity which is resolved using the image.

A main research direction of MMT has been how to best exploit image representations and combine the image and text modalities (Yin et al., 2020; Caglayan et al., 2021; Calixto et al., 2017; Li et al., 2022). It has typically been difficult to surpass strong text-only baselines, the image modality often being ignored (Wu et al., 2021). A major issue holding back progress is that most current state-of-the-art MMT models (Yin et al., 2020; Elliott and Kádár, 2017; Wu et al., 2021; Li et al., 2022) are trained solely on the ∼30k examples of the Multi30k dataset (Elliott et al., 2016), comprising image captions and their translations. This causes two issues: (i) the models do not exploit the large amount of text-only data available and therefore perform poorly in comparison to state-of-the-art text-only MT systems, and (ii) we show that very few examples require images to be correctly translated, which means that the datasets are ill-adapted to evaluating the use of the image modality.

Figure 1: Visual context resolving the ambiguity of English word *glasses* for English-to-French translation.

In this article, we aim to overcome these problems by proposing (i) a new MMT approach that is able to exploit (text-only) monolingual and parallel data as well as (multimodal) captioning data, and that reaches a good balance between maintaining high MT quality and effectively exploiting images, and (ii) a test set, CoMMuTE, containing contrastive evaluation pairs, where images provide the necessary context to disambiguate between multiple meanings of the same source sentence.

Our suggested model is inspired by work on adapting frozen language models (LMs) to multimodal inputs (Sung et al., 2022; Yang et al., 2022; Eichenberg et al., 2021; Pfeiffer et al., 2022); we

[1]https://github.com/MatthieuFP/VGAMT
[2]https://github.com/MatthieuFP/CoMMuTE

propose to adapt a strong MT model to multimodal inputs with lightweight modules (Houlsby et al., 2019) to exploit the large amount of textual data it was trained on. We also propose to better exploit the image by introducing guided self-attention and by combining the standard MMT objective with a visually-conditioned masked language modelling (VMLM) objective (Li et al., 2019; Lu et al., 2019; Su et al., 2020). Our model obtains competitive results compared to strong text-only baselines on standard En→{Fr,De,Cs} MMT benchmarks (Elliott et al., 2016, 2017; Barrault et al., 2018) and outperforms them and state-of-the-art MMT models on our lexically ambiguous contrastive test set.[3]

2 Related Work

Multimodal MT data. The reference dataset to train and evaluate MMT models is Multi30k (Elliott et al., 2016). However, recent work has shown that most MMT systems trained and evaluated on it do not effectively exploit the image information; Elliott (2018) showed that replacing the ground truth image with a random one does not lead to the drop in performance that would be expected, while Wu et al. (2021) argued that the observed gain in performance was due to a regularisation effect. It is also notoriously difficult to beat text-only baselines on this benchmark (Barrault et al., 2018). This may be due to (i) some subsets of Multi30k having been translated independently from the images (Elliott et al., 2016) and (ii) most of the time, the source text being sufficient in theory to produce a perfect translation (i.e. the image is not necessary; see Section 5.2 for our own analysis).

Based on this, alternative test sets and evaluation methods have been proposed. Caglayan et al. (2019) proposed to probe the use of images in MMT models, while Li et al. (2021) proposed another training corpus and evaluation benchmark to evaluate MMT systems, but their work is only based on gender ambiguity and requires specific training data to train MMT models. Lala and Specia (2018) released a lexically ambiguous MMT evaluation dataset to evaluate models ability to disambiguate source sentences, but we found that text context is generally sufficient to translate the evaluation dataset correctly.

[3]CoMMuTE initially contained 50 lexically ambiguous sentences in a previous version of the paper. Results are now computed on the updated version of CoMMuTE comprising 155 ambiguous sentences. Conclusions remain the same.

Contrastive MT datasets. Another means of evaluating (and the one we adopt here) is to target specific phenomena through the use of contrastive test sets. They involve evaluating models based on their ability to rank pairs of translations, where one is correct and the other incorrect. They have been used for the evaluation of different linguistic phenomena, including grammaticality (Sennrich, 2017), multi-sense word disambiguation (Rios Gonzales et al., 2017; Raganato et al., 2019), pronoun translation (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019) and lexical coherence/consistency (Bawden et al., 2018; Voita et al., 2019). Bawden et al. (2018) introduced the idea of conditioning which of the translations is correct depending on linguistic context, and we adopt the same strategy here with our CoMMuTE dataset, composed of lexically ambiguous sentences whose translations are determined by the visual context.

Adapting pretrained LMs to multimodal inputs. A lot of progress has been made through the use of pretrained LMs (Devlin et al., 2019; Conneau and Lample, 2019; Liu et al., 2020), often trained on raw text for text-only models or image captioning data for multimodal ones (Radford et al., 2021; Alayrac et al., 2022; Chen et al., 2022). One of the most efficient ways to learn multimodal LMs is the visually-conditioned masked language modelling (VMLM) objective (Chen et al., 2020; Lu et al., 2019; Su et al., 2020; Li et al., 2020; Zhou et al., 2021; Huang et al., 2021a; Li et al., 2019). Inspired by the masked language modelling (MLM) objective (Devlin et al., 2019), it consists in randomly masking input text tokens and predicting them conditionally based on the visual features. A lot of interest has also been shown in lightweight modules such as adapters (Houlsby et al., 2019) to adapt large frozen LMs to multimodal tasks (Eichenberg et al., 2021; Yang et al., 2022; Pfeiffer et al., 2022; Tsimpoukelli et al., 2021; Sung et al., 2022) in order to avoid catastrophic forgetting (De Lange et al., 2021). Based on these approaches, we propose to adapt a strong text-only MT model with lightweight modules in order to exploit the large amount of data it previously learned.

Which type of visual features in MMT systems? In terms of how images are represented in multimodal models, different strategies exist. Many works first proposed to incorporate global visual features from object recognition models pretrained

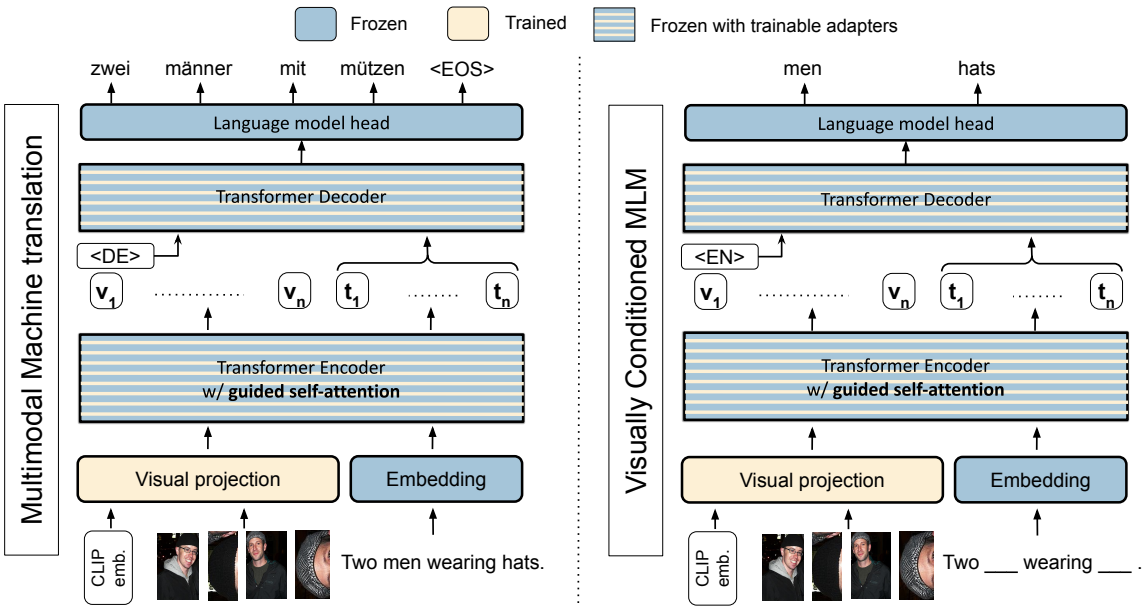


Figure 2: Overview of our approach, multimodal MT (MMT) (left) and visually-conditioned masked language modeling (VMLM) (right) objectives. We train VGAMT on both objectives jointly.

on ImageNet (Deng et al., 2009), such as ResNet50 (He et al., 2016), either in the form of a single vector or a set of features (Calixto et al., 2017; Elliott and Kádár, 2017; Calixto and Liu, 2017; Yao and Wan, 2020; Helcl et al., 2018). More recent global features extractor such as CLIP (Radford et al., 2021) exist, but to our knowledge have not been used in MMT models. Extending this idea, other works focused on entities in the image and extracted bounding boxes using a pretrained Faster R-CNN (Ren et al., 2015) in order to introduce more semantic visual information into MT (Grönroos et al., 2018; Ive et al., 2019; Caglayan et al., 2021). Recent efforts have been made to only select parts of the image that are relevant to the translation of the sentence. Some proposed to use a more selective attention mechanism between modalities (Liu et al., 2021; Ye et al., 2022), while others suggested extracting other types of visual features (Huang et al., 2021b; Fang and Feng, 2022). Based on this, Yin et al. (2020) decided to exploit local image-text correspondences in their model Graph-MMT. Similar to their approach, we use a simpler method to extract relevant visual features, using the output queries from a state-of-the-art free-form text object detector MDETR (Kamath et al., 2021) as our local visual features (in addition to global features from CLIP).

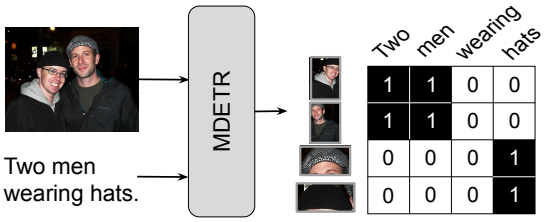


Figure 3: Example of an MDETR alignment matrix.

3 Our approach: VGAMT

The two main aims of our approach are to (i) exploit a maximum available data (not just multimodal parallel text data) and to (ii) provide an effective way to combine image and text modalities. Our approach, shown in Figure 2, consists in taking a strong text-only MT model[4] and adapting it to multimodal MT. To adapt this strong text-only model to multimodal inputs, we add several lightweight modules—bottleneck adapters (Houlsby et al., 2019) and linear visual projection layers—to the otherwise frozen initial model. The bottleneck adapters are lightweight linear layers introduced after each attention block and each feed-forward layer to project embeddings down before projecting them up.

In terms of representing visual information, we choose to use two types of representation. We concatenate local (MDETR) features and global

[4]In practice, our starting point is mBART, which we fine-tune on a large parallel corpus (see Section 5 for more details).

(CLIP) features to the text inputs. We choose to use global features too, since the source sentence can describe more general aspects of the image than mere objects (such as scenes). We jointly train the non-frozen parts of our model on two distinct objectives: multimodal MT (MMT) and visually-conditioned masked language modelling (VMLM), as described in Section 3.1. We also introduce a guided self-attention to exploit image information in a straightforward manner (see Section 3.2) in the encoder (while the decoder uses regular self- and cross-attentions and can only attend to embeddings related to text positions). We call our approach Visually Guided and Adapted Machine Translation (VGAMT).

3.1 Combining training objectives

As shown in Figure 2, we jointly train VGAMT on two objectives: visual masked language modelling (VMLM) and multimodal MT (MMT). VMLM (resp. MMT) consists in predicting masked tokens (resp. translating the sentence) conditioned on the image.[5] The use of the VMLM objective in addition to MMT ensures that the model does not learn to ignore the visual inputs when translating (since Multi30k is mainly composed of very standard and unambiguous parallel sentences). We make sure to mask a high percentage (25%) of the text inputs so that the model is forced to attend to the image when producing translations.

3.2 Guided self-attention

The backbone of VGAMT is an encoder-decoder MT model, in which image features are concatenated to textual input embeddings and shared self-attention is used over the two input modalities (see Figure 2). Instead of using full self-attention (Caglayan et al., 2021) (connections between all image parts and all text tokens), we introduce guided self-attention. Guided self-attention consists in masking irrelevant connections between text and image representations; each text (resp. image) embedding can attend to itself and all other text (resp. image) positions, but can only attend to image (resp. text) positions conditioned on pre-extracted text-image alignments. We obtain these alignments (in the form of a cross-modal correspondence matrix) using MDETR (Kamath et al., 2021), which detects image regions and corresponding text spans based on a free-form text (see Figure 3 and Appendix B for more details).

Concretely, let Q, K and V denote the learnable query, key and value parameters of a standard self-attention mechanism. Attention can be defined as $\text{Attention}(Q, K, V) = A \cdot V$, where the attention matrix $A = (a_{ij})$ is defined as $A = \text{softmax}\left(QK^T/\sqrt{d_k}\right)$, where d_k is the dimension of the key vector, i.e.:

$$a_{ij} = \frac{e^{Q_i K_j^T/\sqrt{d_k}}}{\sum_l e^{Q_i K_l^T/\sqrt{d_k}}} \quad (1)$$

The idea behind our guided self-attention mechanism is that we want to allow subwords to attend to all subwords, all bounding boxes to attend to all bounding boxes, but to only allow cross-modal attention between a subword and bounding boxes that are linked by MDETR (see Figure 3). We therefore define a binary masking matrix $C = (c_{ij})$ where (i) $c_{ij} = 1$ if indices i and j correspond to embeddings coming from the same modality, and (ii) c_{ij} is provided by the MDETR matrix otherwise: it is 1 if MDETR has created a link between subword (resp. bounding box) i and bounding box (resp. subword) j. Once this *guiding* matrix C is defined, we can replace the standard attention (1) with our guided attention:

$$a_{ij} = \frac{c_{ij} e^{Q_i K_j^T/\sqrt{d_k}}}{\sum_l c_{il} e^{Q_i K_l^T/\sqrt{d_k}}}. \quad (2)$$

The main advantage of guided self-attention over full self-attention is that the model does not have to learn to ignore irrelevant text-image correspondences since alignments are introduced as a prior.

4 Contrastive Multilingual Multimodal Translation Evaluation (CoMMuTE)

To overcome the flaws of existing benchmarks (see Section 5.2), we introduce CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation dataset[6]. It is composed of 155 lexically ambiguous sentences in English, each associated with two translations corresponding to two of the possible meanings of each sentence and two images that determine which of the translations is correct. It covers English→French, English→German and English→Czech. An example is given in Figure 4.

[5] During training, we randomly draw batches from a parallel multimodal dataset (for MMT) and a monolingual multimodal one (for VMLM) with equal probability.

[6] CoMMuTE is distributed under Creative Commons Attribution Share Alike 4.0 International license.

Data collection. The test set contains 155 ambiguous sentences constructed around 155 lexically ambiguous words: 29 of the examples are from Bawden et al. (2018), and we created the remaining ones.[7] We collected two images for each sentence under Creative Commons license (either Google Images or our own photos), so that the image illustrates without ambiguity one of the two meanings of the sentence. We do not restrict the image-text relation to be strictly descriptive (as for image captions) in order to have a more general evaluation dataset. Each sentence was translated into two possible translations (each corresponding to one of the images) by a native speaker of the target language. Appendix A provides some basic statistics.

The idea of CoMMuTE is to use MMT models to rank each of the two translations based on image information. The perplexity of a sentence for a given model is defined as: $PPL_q(y) = \prod_{i=1}^{N} q(y_i)^{-\frac{1}{N}}$, where q is the probability distribution output by the model, N is the sequence length and $y_1, \ldots, y_N$ is the sequence of tokens. Now, let $y_1, \ldots, y_{N_1}$ be the sequence of tokens of the correct translation and $y'_1, \ldots, y'_{N_2}$ the sequence of tokens of the incorrect translation, a model makes a correct prediction if $PPL_q(y) \leq PPL_q(y')$. i.e. the model considers the correct translation more likely than the incorrect one. For each example, we rank each of the translations based on each of the images (2 comparisons per example), and report the accuracy over all the examples. As CoMMuTE is perfectly balanced, a text-only model will get exactly 50% accuracy on this task.

	En→Fr		En→De		En→Cs	
	size	#sents.	size	#sents.	size	#sents.
OpenSubtitles	2.2GB	24.2M	1.2GB	13.1M	2.2GB	24.7M
Ted Talks	108MB	535K	83MB	414K	30MB	158K
Books	29MB	119K	12MB	47K	-	
Wikipedia	187MB	769K	493MB	2.2M	3.2MB	19K
Total	2.5GB	25.6M	1.8GB	15.8M	2.2GB	24.9M

Table 1: Parallel corpus sizes.

5 Experiments

5.1 Text-only data

All our experiments are based on the strong MT model mBART[8] (Liu et al., 2020), which we fine-tune on parallel text (see Table 1). We use OpenSubtitles2018[9] (Lison et al., 2018), Wikipedia (Wołk and Marasek, 2014), Ted Talks (Reimers and Gurevych, 2020) and the Books datasets (Tiedemann, 2012). We preprocess the data using Moses scripts (Koehn et al., 2007).[10]

Figure 4: Example from CoMMuTE. The English word 'mole' can refer either to 'a small dark, raised lump on a the skin' (1) or 'a small burrowing mammal' (2).

5.2 Multimodal data

	Test2016	Test2017	MSCOCO
Ambiguous (%)	21 (2.1%)	20 (2%)	6 (1.3%)

Table 2: Number (and percentage) of ambiguous examples in the En→Fr test sets.

Multi30k. We train our frozen MT model on the Multi30k dataset (Specia et al., 2016; Elliott et al., 2016) composed of English sentences, each accompanied by an image and French, German and Czech translations. It contains 29k train, 1014 dev and 1000 test examples (Test2016). Elliott et al. (2017) and Barrault et al. (2018) released two additional related test sets (Test2017 and Ambiguous Coco). However, on analysis of these sets and as shown in Table 2, we found that very few examples are image-dependent (i.e. the source sentence is ambiguous and the image is required to solve the ambiguity in the target language),[11] meaning that an MMT system is unlikely to perform better than a text-only system. Moreover, most of these ambiguities are semantically similar and they only cover a few multi-sense words. Although Ambiguous Coco (Elliott et al., 2017) is designed to be an

[7] We could not take the entirety of the examples in (Bawden et al., 2018) as some examples were not adapted to disambiguation using visual (as opposed to linguistic) context.

[8] mBART is pretrained on CC25 (Wenzek et al., 2020).

[9] http://www.opensubtitles.org

[10] `remove-non-printing-char.pl`, `normalization-punctuation.pl`, and `clean-corpus-n.pl` (4-100 tokens).

[11] We queried WordNet (Fellbaum, 1998) for all nouns and verbs in the English sentences. An example was considered image-dependent if, on manual assessment, there were multiple meanings, which were (i) compatible with the text context and (ii) could be disambiguated using the image.

ambiguous test set as it is built around multi-sense verbs, it was automatically created from sentences from MSCOCO (Lin et al., 2014) for which the textual context is often sufficient for disambiguation. These benchmarks remain useful to make sure MMT systems do not perform worse than text-only MT models on examples where images are not necessary to translate correctly. However, we consider them insufficient to assess how well MMT systems exploit images to improve translation.

Monolingual multimodal data. For the VMLM objective, we train our model on the Conceptual Captions (CC) dataset (Sharma et al., 2018) composed of 3.3M[12] images aligned with English text.

5.3 Implementation details

For all our experiments, we use the mBART implementation from Hugging Face (Wolf et al., 2020). Experiments with adapters used bottleneck adapters (Houlsby et al., 2019) with a reduction factor of 8 and ReLU activation (Agarap, 2018). We use the implementation provided by adapter-transformers (Pfeiffer et al., 2020). We use a batch size of 512, the Adam optimiser (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and a learning rate of 10^{-4} for En→Fr and 10^{-5} for En→{De,Cs}. We also applied 0.1 label smoothing (Szegedy et al., 2016) during training. We selected our final model according to the best BLEU score (Papineni et al., 2002) on the Multi30k dev set after at least one full pass over the Multi30k and Conceptual Captions training sets. We ran each experiment 3 times with different seeds and report the average BLEU[13] (Papineni et al., 2002) and COMET (Rei et al., 2020) scores[14] and the standard errors. We also report METEOR scores (Banerjee and Lavie, 2005) in Appendix E. All experiments were carried out on 8 NVIDIA V100 GPUs for ∼15h.

5.4 Baselines

We consider several text-only and multimodal baselines. All baselines except the MT models fine-tuned from mBART were trained from scratch with the original codebases and features released by the papers' authors. Models trained on the (multimodal) MT objective only where trained on Multi30k, while models jointly trained on the (multimodal) MT and (V)MLM objectives were trained on Multi30k and Conceptual Captions.

Text-only. We trained a text-only Seq2seq Transformer (Vaswani et al., 2017) from scratch and a text-only Seq2Seq Transformer initialised from TLM weights (Conneau and Lample, 2019). We refer to these models as Vanilla MT and TLM + MT respectively. We also trained several MT models initialised from pretrained mBART (Liu et al., 2020) and which we fine-tuned on parallel data (Lison et al., 2018; Wołk and Marasek, 2014). We refer to these models as mBART + MT. '*w/ adapters*' specifies that the model's weights are frozen except bottleneck adapters (Houlsby et al., 2019).

Multimodal. We trained several state-of-the-art multimodal MT models: Graph-MMT (Yin et al., 2020), Gated Fusion (Wu et al., 2021) and a Seq2Seq Transformer trained from VTLM weights (Caglayan et al., 2021) (hereafter VTLM + MMT).

5.5 Results and Analysis

Figure 5: Example from Test2017 with English *biker*, translated as French 'cycliste' (cyclist) or 'motard' (motorcyclist). VGAMT succeeds where the baseline fails.

Tables 3 and 4 show BLEU, COMET and accuracy scores for all models compared on several En→{Fr,De,Cs} test sets including CoMMuTE. An initial observation is that the text-only model is a strong baseline on the three standard benchmarks (Test2016, Test2017 and MSCOCO). As mentioned in Section 5.2, most of these evaluation datasets do not need visual context to be correctly translated. Our model VGAMT is on average on par with its counterpart text-only mBART+MT *w/ adapters* baseline for all Multi30k En→Fr test sets, while being on average just below this baseline on En→{De,Cs} Multi30k benchmarks. It outperforms other MMT models with a large margin due to both the effective use of textual knowledge from

[12]At the time of writing, we were able to collect ∼2M images and trained models on this subset.

[13]computed with the default parameters of sacreBLEU v2.0.0 from https://github.com/mjpost/sacrebleu.

[14]Using the `wmt20-comet-da` model.

Model	Objectives	# trainable params	Test2016 BLEU	Test2016 COMET	Test2017 BLEU	Test2017 COMET	MSCOCO BLEU	MSCOCO COMET	CoMMuTE Accuracy
En→Fr									
Text-only Machine Translation									
Vanilla MT*	NMT	4.0M	59.4 ±0.2	0.711 ±0.004	51.6 ±0.2	0.568 ±0.009	41.2 ±0.4	0.403 ±0.005	*50.0*
TLM + MT*	NMT	42M	62.0 ±0.1	0.795 ±0.002	54.2 ±0.2	0.681 ±0.002	43.6 ±0.2	0.542 ±0.009	*50.0*
mBART + MT*	-	-	49.0	0.819	48.1	0.779	47.0	0.733	*50.0*
mBART + MT* *w/ adapters*	NMT + MLM	12.6M	**67.2** ±0.3	**0.971** ±0.005	61.5 ±0.3	0.918 ±0.004	**51.5** ±0.7	**0.832** ±0.006	*50.0*
Multimodal Machine Translation									
Graph-MMT*	MMT	4.0M	58.9 ±0.5	0.705 ±0.004	51.5 ±0.2	0.589 ±0.005	41.0 ±0.6	0.387 ±0.013	50.2 ±3.5
Gated Fusion*	MMT	2.8M	58.7 ±0.3	0.707 ±0.002	50.8 ±0.7	0.580 ±0.011	40.4 ±0.4	0.394 ±0.013	50.0 ±0.8
VTLM + MMT*	MMT	44M	61.4 ±0.2	0.783 ±0.005	53.6 ±0.1	0.672 ±0.005	43.4 ±0.3	0.500 ±0.006	50.1 ±0.3
VGAMT (*ours*)	MMT + VMLM	13.2M	**67.2** ±0.1	0.968 ±0.002	**61.6** ±0.1	**0.921** ±0.002	51.1 ±0.6	0.811 ±0.003	**67.1** ±0.7
En→De									
Text-only Machine Translation									
Vanilla MT*	NMT	4.1M	38.5 ±0.3	0.394 ±0.005	30.3 ±0.5	0.259 ±0.012	27.8 ±0.4	0.092 ±0.018	*50.0*
TLM + MT*	NMT	42M	40.0 ±0.2	0.457 ±0.006	31.5 ±0.1	0.341 ±0.002	29.4 ±0.3	0.152 ±0.015	*50.0*
mBART + MT*	-	-	36.2	0.595	32.3	0.506	27.6	0.383	*50.0*
mBART + MT* *w/ adapters*	NMT + MLM	12.6M	**43.6** ±0.2	**0.697** ±0.003	**38.9** ±0.5	**0.664** ±0.002	**36.2** ±0.2	**0.574** ±0.004	*50.0*
Multimodal Machine Translation									
Graph-MMT*	MMT	4.1M	38.6 ±0.3	0.368 ±0.011	29.0 ±0.5	0.226 ±0.010	25.9 ±0.8	0.060 ±0.027	49.1 ±1.5
Gated Fusion*	MMT	2.8M	38.7 ±0.2	0.378 ±0.007	29.5 ±0.2	0.236 ±0.018	26.6 ±0.3	0.055 ±0.016	49.7 ±0.6
VTLM + MMT*	MMT	44M	39.4 ±0.2	0.439 ±0.004	30.7 ±0.2	0.322 ±0.005	28.2 ±0.2	0.168 ±0.014	50.0 ±0.2
VGAMT (*ours*)	MMT + VMLM	13.2M	43.3 ±0.2	0.694 ±0.003	38.3 ±0.2	0.653 ±0.005	35.7 ±0.3	0.544 ±0.006	**59.0** ±0.5

Table 3: Results for En→{Fr,De} (average of three runs). The best result is indicated in **bold**. * means the results were retrained by using the original codebase provided by the authors of the paper.

the frozen MT model but also guided self-attention. Note that the scores reported for the baselines are lower than the ones reported in the original papers of the models for several reasons. First, we computed the scores on fully detokenised data to have a uniform evaluation between all models. We also report the average score from three different runs using different seeds and not the best score obtained over a single run.

More importantly, our VGAMT obtains strong improvements over both text-only baselines and state-of-the-art MMT systems on CoMMuTE; our model can use visual context to disambiguate sentences. This can be seen in Figure 5 (one of the ambiguous examples from Multi30k), where in contrast to the baseline VGAMT produces the correct translation and Figure 6 (from CoMMuTE), where VGAMT correctly ranks the two translations. More examples are provided in Appendix D. We also propose to translate CoMMuTE source sentences and compare against the reference translations; the results are shown in Appendix F.

Figure 6: VGAMT Perplexity scores on a CoMMuTE example, illustrating that it is able to correctly rank each of the French translations of ambiguous English *bucks* 'male deer or dollars' when conditioning on the image.

Model	Test2016 BLEU	Test2016 COMET	Test2018 BLEU	Test2018 COMET	CoMMuTE Accuracy
En→Cs					
Text-only Machine Translation					
Vanilla MT*	31.3 ±0.0	0.593 ±0.008	26.0 ±0.2	0.379 ±0.008	*50.0*
TLM + MT*	32.6 ±0.1	0.642 ±0.002	26.8 ±0.2	0.432 ±0.006	*50.0*
mBART + MT*	32.1	0.865	29.6	0.747	*50.0*
w/ adapters	37.3 ±0.1	**0.940** ±0.005	**35.2** ±0.4	**0.876** ±0.002	*50.0*
Multimodal Machine Translation					
Graph-MMT*	30.8 ±0.4	0.562 ±0.011	24.9 ±0.5	0.344 ±0.011	49.2 ±1.8
Gated Fusion*	30.8 ±0.4	0.560 ±0.014	25.8 ±0.1	0.342 ±0.008	51.0 ±1.9
VTLM + MMT*	32.0 ±0.3	0.621 ±0.010	26.7 ±0.2	0.419 ±0.015	50.0 ±0.3
VGAMT (*ours*)	**37.6** ±0.2	0.934 ±0.004	34.2 ±0.1	0.833 ±0.003	**55.6** ±0.8

Table 4: Results for En→Cs (average of three runs). Same formatting as for Table 3.

6 Ablation Study

To better understand the role of VGAMT's components, we carry out several ablations for En→Fr and report all results in Table 5.

Adapters versus Fine-tuning. We compare the results of fine-tuning an unfrozen VGAMT model (w/o adapters) in comparison to our frozen model with adapters (VGAMT), all other things remaining equal. The unfrozen version faces a drop in

Model	Test2016 BLEU	Test2016 COMET	Test2017 BLEU	Test2017 COMET	MSCOCO BLEU	MSCOCO COMET	CoMMuTE Accuracy
Text-only Machine Translation							
mBART + MT* *w/ adapters*	67.2 ±0.3	0.971 ±0.005	61.5 ±0.3	0.918 ±0.004	**51.5** ±0.7	**0.832** ±0.006	*50.0*
w/o MLM objective	**67.7** ±0.3	0.970 ±0.004	61.5 ±0.1	**0.926** ±0.004	50.3 ±0.4	0.821 ±0.002	*50.0*
Multimodal Machine Translation							
VGAMT (*ours*)	67.2 ±0.1	0.968 ±0.002	**61.6** ±0.1	0.921 ±0.002	51.1 ±0.6	0.811 ±0.003	**67.1** ±0.7
unfrozen w/o adapters	66.9 ±0.7	0.965 ±0.003	61.4 ±0.6	0.912 ±0.009	50.3 ±0.7	0.814 ±0.011	60.5 ±3.8
w/o VMLM objective	**67.7** ±0.2	**0.976** ±0.001	61.4 ±0.2	0.920 ±0.003	50.5 ±0.0	0.809 ±0.004	52.0 ±1.2
w/o guided self-attention	67.0 ±0.2	0.963 ±0.004	60.8 ±0.3	0.910 ±0.006	50.3 ±0.5	0.792 ±0.004	64.6 ±1.6
w/ pretraining (w/o co-training)	66.2 ±0.1	0.950 ±0.001	59.3 ±0.1	0.875 ±0.003	49.2 ±0.2	0.777 ±0.001	63.3 ±0.5
w/o MDETR features	66.7 ±0.5	0.967 ±0.004	61.1 ±0.1	0.912 ±0.002	51.0 ±0.6	0.810 ±0.003	63.0 ±1.2
w/o CLIP features	66.4 ±0.8	0.959 ±0.008	60.4 ±0.7	0.909 ±0.002	51.0 ±0.6	0.810 ±0.008	50.3 ±0.0

Table 5: Results of the ablation studies described in Section 6 (En→Fr test set). The best result is indicated in **bold**.

scores on all test sets except Test2017. Notably, the unfrozen model's accuracy score of 60.5 on CoMMuTE is 6.6 points lower than our final VGAMT model. As well as providing a more lightweight solution that does not involve fine-tuning all parameters, using neural adapters and freezing other weights is useful in terms of performance.

Impact of the VMLM objective. To evaluate the impact of jointly training with MMT and VMLM objectives, we train a model on the MMT without VMLM (and therefore without monolingual multimodal data). The MMT model trained on MMT alone obtains 52.0 on CoMMuTE, compared to 67.1 for joint training, showing that VMLM helps our model to better exploit disambiguating images.

Guided self-attention. We study the impact of guided self-attention between modalities by comparing against classic full self-attention. Guided self-attention obtains better results than full self-attention, particularly on Test2017 and MSCOCO (+0.8 BLEU, +0.015 COMET on average). It also gets better results on CoMMuTE (+2.5 points). See Appendix C for analysis of guided attention scores.

VMLM and MMT joint training. We compare our VMLM and MMT joint training with disjoint training where VGAMT is first pretrained on VMLM then fine-tuned on MMT instead of co-training on both VMLM and MMT. Table 5 shows that it results in a large drop of performance on all scores in average including 3.8 points on CoMMuTE.

MDETR. We examine the impact of MDETR features by training a model without them.[15] The results without MDETR features are slightly lower than the full model on standard MMT benchmarks. However, the results are significantly lower on CoMMuTE (63.0±1.2 without MDETR features and 67.1±0.7 with MDETR features). This means that VGAMT benefits from MDETR features when disambiguating and translating sentences.

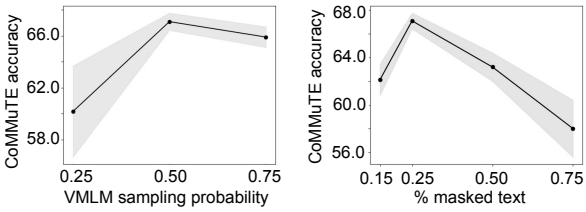
CLIP. We also study the impact of CLIP features by training a model without them.[15] Including CLIP features gives slightly higher results on standard MMT benchmarks (+0.69 BLEU and +0.007 COMET scores on average on all benchmarks). VGAMT without CLIP features faces an extreme drop on CoMMuTE (50.3±0.00 w/o CLIP features vs. 67.1±0.7 w/ CLIP features), which shows that CLIP features are required for disambiguation.

VMLM sampling probability and degree of masking. We ran experiments to vary the VMLM sampling probability (see Section 3.1) and the percentage of masked text inputs (see Figure 7 for results on CoMMuTE). For the sampling between VMLM and MMT objectives, the maximum value is reached for p =50%, i.e. equal sampling between VMLM and MMT objectives (Figure 7a). Similar results are obtained for $p = 75\%$, i.e. 3 VMLM batches for 1 MMT batch, but the translation quality is lower. For the percentage of masking, there is a peak at 25% masked text inputs and a constant decrease for higher values (Figure 7b).

7 Conclusion

We propose a new MMT approach (VGAMT) based on (i) adapting a strong text-only MT model with lightweight adapters and (ii) introducing better use of the text and image modalities through a novel guided self-attention mechanism and joint MMT and VMLM training. We also introduce

[15]More details are available in Appendix B.



(a) Variation of VMLM sampling probability with fixed 25% masked text inputs.

(b) Variation of masked text inputs with fixed 50% VMLM sampling probability.

Figure 7: CoMMuTE Results comparing multiple VMLM sampling probabilities and percentage of masked text inputs. 95% confidence interval in grey.

CoMMuTE, a contrastive test set designed to test the use of visual disambiguating context. Results for En→{Fr,De,Cs} show that VGAMT obtains competitive results compared with strong text-only baselines on standard benchmarks and widely outperforms these baselines and state-of-the-art MMT systems on CoMMuTE.

Limitations

In this work, we focused on En→{Fr,De,Cs} multimodal MT. At the time of writing, our method can only be applied for En→X MMT. It is indeed necessary to have access to a modulated object detector in the source language to extract the features and the image-text relationship exploited by our model. This type of modulated object detector is only available in English for the moment. We leave the extension of our method to non-English source languages to future work. Moreover, our method requires large amount of captioning data to perform well. It is therefore computationally expensive.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013908 and 2022-AD011012254 made by GENCI. It was also partly funded by the last four authors' chairs in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

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.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. 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, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

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

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian

Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

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 *European conference on computer vision*, pages 104–120. Springer.

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

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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.

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

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*.

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, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

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

Christiane Fellbaum. 1998. *WordNet 1.6: An Electronic Lexical Database*. Bradford Books. MIT Press.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Kaiming He, X. 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.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

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 *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Haoyang Huang, Lin Su, Di Qi, Nan Duan, Edward Cui, Taroon Bharti, Lei Zhang, Lijuan Wang, Jianfeng Gao, Bei Liu, Jianlong Fu, Dongdong Zhang, Xin Liu, and Ming Zhou. 2021a. M3P: Learning universal representations via multitask multilingual

multimodal pre-training. *In 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3976–3985.

Xin Huang, Jiajun Zhang, and Chengqing Zong. 2021b. Entity-level cross-modal learning improves multimodal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1067–1080, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multimodal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu (Richard) Chen, Rogerio Feris, David Cox, and Nuno Vasconcelos. 2022. VALHALLA: Visual Hallucination for Machine Translation. In *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5226.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2021. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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.

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.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. 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, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

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.

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

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

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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5227.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Krzysztof Wołk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, 18:126–132. International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014, Warsaw, Poland.

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*.

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

Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022. Noise-robust cross-modal interactive learning with Text2Image mask for multimodal neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5098–5108, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC^2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, Nashville, TN, USA.

A CoMMuTE statistics

Some basic statistics of the CoMMuTE dataset can be found in Table 6. The source side of the dataset is always English and two translations of each of the 155 English ambiguous sentences are provided in French, German and Czech.

	En	Fr	De	Cs
#unique sents.	155	308	300	308
Avg. sent. length	6.54	6.90	6.48	5.07
#unique toks	462	679	638	718

Table 6: CoMMuTE statistics.

B Visual features

We use MDETR (Kamath et al., 2021) features as our local visual features. Concretely, we extract the set of output queries features of size 64 from the MDETR decoder and introduce them as input.

In addition, we use CLIP (Radford et al., 2021) features as our global visual features. More specifically, we extract the output [CLS] features of size 512 from the ViT (Dosovitskiy et al., 2021) image encoder used by CLIP and introduced it as input.

C Guided self-attention analysis

We studied the values of the cross-modal part of our guided self-attention. To do so, we followed the method proposed by Kobayashi et al. (2020) who showed that raw attention scores α are meaningless and instead proposed to conduct analysis on the normalised attention scores $\|\alpha f\|$, where α are the raw attention scores and f is the value vector in the attention mechanism. Figure 9b shows the cross-modal part of the guided self-attention map from the example displayed in Figure 9a where all the values have been averaged over all heads and all layers. In this example, the English word *fans* 'cooling device or ardent admirer' is ambiguous and the two meanings have different translations in

	En → Fr			En → De			En → Cs	
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	Test2016	Test2018
Text-only Machine Translation								
Vanilla MT*	74.4 ±0.1	68.3 ±0.2	61.5 ±0.4	55.0 ±0.2	46.9 ±0.4	45.3 ±0.3	30.5 ±0.1	26.5 ±0.1
TLM + MT*	76.3 ±0.1	70.3 ±0.2	63.4 ±0.3	56.0 ±0.2	48.1 ±0.1	46.1 ±0.2	31.0 ±0.0	26.6 ±0.1
mBART + MT*	68.3	66.8	66.4	52.6	48.3	44.2	30.7	28.1
mBART + MT* *w/ adapters*	**79.9** ±0.3	**76.0** ±0.2	**69.5** ±0.6	**58.5** ±0.1	**53.9** ±0.3	**51.7** ±0.2	**33.8** ±0.2	**31.4** ±0.2
Multimodal Machine Translation								
Graph-MMT*	74.1 ±0.4	68.7 ±0.5	61.6 ±0.6	54.4 ±0.4	45.7 ±0.4	43.2 ±0.7	30.1 ±0.1	26.0 ±0.2
Gated Fusion*	73.1 ±0.3	67.1 ±0.5	60.1 ±0.4	54.9 ±0.4	46.2 ±0.3	44.2 ±0.4	28.8 ±0.2	25.1 ±0.1
VTLM + MMT*	75.9 ±0.1	69.8 ±0.1	63.3 ±0.2	55.4 ±0.1	47.7 ±0.1	45.6 ±0.3	30.6 ±0.1	26.4 ±0.1
VGAMT (*ours*)	79.7 ±0.0	75.9 ±0.1	68.9 ±0.4	58.1 ±0.2	53.6 ±0.2	**51.7** ±0.2	33.7 ±0.1	30.5 ±0.0

Table 7: METEOR scores for standard En→{Fr,De,Cs} benchmarks (average of three runs). The best result is indicated in **Bold**. * means the results were retrained by using the original codebase provided by the authors of the paper.

French, German and Czech. Given the region-text couples extracted by MDETR (Figure 9a), only the token *fans* can attend to the MDETR region features. The normalised attention scores of the embedding of the token *fans* on these regions are low in comparison to the scores on the text part and on the CLIP embedding. On the contrary, all embeddings can attend to CLIP embedding and the embedding of the token *fans* is the one with the highest normalised attention score with CLIP embedding.

D Additional examples

EN Two men in uniforms are playing **football** in the snow.

FR Deux hommes en tenues jouent au **football américain** dans la neige.

Text-only Deux hommes en tenues jouent au **football** dans la neige.

VGAMT Deux hommes en tenues jouent au **football américain** dans la neige.

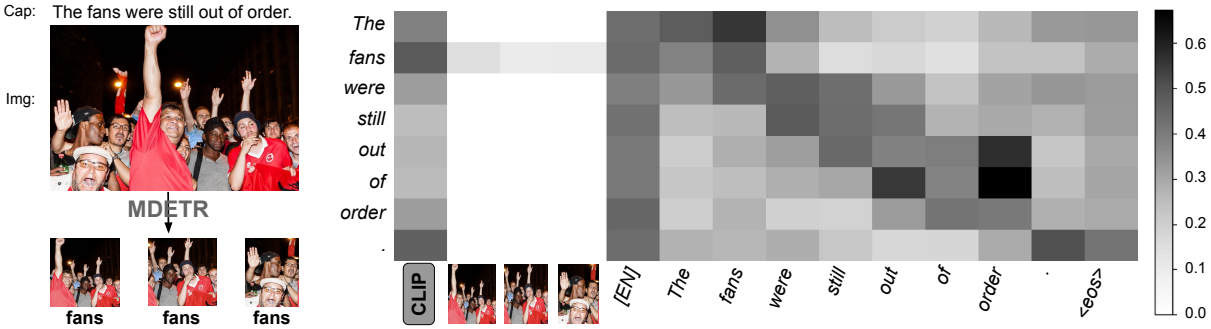
Figure 8: Example from Test2017, illustrating how VGAMT is able to exploit visual information to distinguish between the two types of football (soccer (1) and American football (2) depending on whether British or American English is used), whereas the text-only baseline produces a wrong translation.

Figure 12 shows examples from CoMMuTE and the perplexity scores obtained by VGAMT. It is able to choose the correct translations from English sentences with the ambiguous words *chips*, *bugs*, *red light*. However, it fails to choose the correct translation in the first case of Figure 12d; the picture shows a beam 'ray of light' and the perplexity of the correct (top) translation with the French translation *rayon* is higher than the incorrect (bottom) one with the French translation *poutre*. Nevertheless, the model gives a lower perplexity to the sentence with the correct image (1.847) in comparison to the same sentence with the incorrect image (2.616). So, even if VGAMT is not able to choose the correct translation in the first case of this example, it shows some evidence of being able to discriminate between the French translation with the correct image and the same French translation with the incorrect image. Figures 12e and 12f show two other similar examples in En→De MT.

In terms of translation (rather than reranking), Figure 8 shows an example from Multi30k where our model correctly translates the ambiguous word while the text-only baseline fails to do so.

Model	Test2016	Test2017	MSCOCO
Text-only Machine Translation			
mBART + MT* *w/ adapters*	79.9 ±0.3	76.0 ±0.2	**69.5** ±0.6
w/o MLM objective	**80.3** ±0.2	**76.3** ±0.2	68.7 ±0.3
Multimodal Machine Translation			
VGAMT (*ours*)	79.7 ±0.0	75.9 ±0.1	68.9 ±0.4
unfrozen w/o adapters	79.8 ±0.5	75.8 ±0.2	68.7 ±0.6
w/o VMLM objective	80.3 ±0.1	76.0 ±0.1	68.7 ±0.1
w/o guided self-attention	79.6 ±0.1	75.4 ±0.2	68.4 ±0.3
w/ pretraining (w/o co-training)	79.2 ±0.0	74.3 ±0.1	67.9 ±0.2
w/o MDETR features	79.5 ±0.3	75.6 ±0.1	68.9 ±0.6
w/o CLIP features	79.2 ±0.5	75.2 ±0.3	69.0 ±0.5

Table 8: METEOR scores for the ablations described in Section 6 (En→Fr). The best result is indicated in **bold**.



(a) English sentence, its associated image and the region-text couples extracted by MDETR.

(b) Normalised attention scores from the cross-modal part of our guided self-attention mechanism. Only the embeddings of the token '*fans*' can attend to MDETR features. All the embeddings can attend to CLIP features.

Figure 9: Guided self-attention map for the English sentence *The fans were still out of order* and its associated image. The sentence is ambiguous as English *fan* refers to 'a cooling device' or 'an ardent admirer'. Values are averaged over all heads and all layers. <*eos*> refers to the end of sentence token.

E METEOR scores

In order to compare to previous work, we also provide METEOR scores in Table 7 for En→{Fr,De,Cs} standard benchmarks. It confirms that VGAMT obtains competitive results over a strong text-only baseline on benchmarks where images are not necessary for translation. METEOR scores for the En→Fr ablations conducted in Section 6 are shown in Table 8.

F Translating CoMMuTE

		VGAMT (*ours*)	mBART + MT* *w/ adapters*
En→Fr	BLEU	32.2 ±1.7	**34.5** ±1.4
	COMET	**0.362** ±0.048	0.306 ±0.014
	METEOR	48.5 ±2.1	**52.3** ±1.4
En→De	BLEU	**29.3** ±0.6	25.9 ±0.7
	COMET	**0.184** ±0.024	0.182 ±0.007
	METEOR	**43.0** ±0.8	41.3 ±1.3
En→Cs	BLEU	**20.8** ±0.9	18.3 ±1.3
	COMET	**0.525** ±0.024	0.491 ±0.022
	METEOR	**23.4** ±0.8	22.4 ±0.7

Table 9: MT Generation results for CoMMuTE. Best results are indicated in **bold**.

CoMMuTE is designed as a contrastive test set to be used for reranking. However, it is possible to translate the source sentences too and compare against the reference translations.

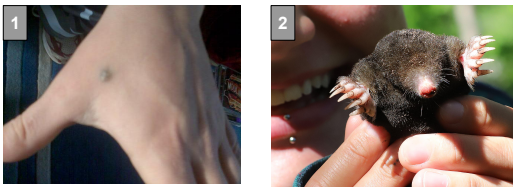
Table 9 shows the MT results on CoMMuTE comparing VGAMT and the strong text-only baseline. They may indicate that traditional metrics for MT task are ill-adapted to evaluating the use of visual information by MMT models. For instance, BLEU and METEOR scores for the text-only baseline are significantly higher than the scores for our model VGAMT on the En→Fr split whereas our

Figure 10: Machine Translation example from CoMMuTE. VGAMT is able to exploit visual information to translate '*bow*' correctly in the two cases.

VGAMT obtains 67.10 accuracy on the contrastive evaluation (Table 3). It might be due to the fact that such metrics are less reliable on small datasets or that BLEU and METEOR are words matching metrics and therefore output low scores for synonyms or similar content described differently. On the other hand, COMET is an embedding-based metric, which outputs higher scores for synonyms which may be why VGAMT outperforms the text-only baseline with this metric; as illustrated by Figure 10 where VGAMT outputs *noeud* 'bow' which is a synonym of the reference translation *ruban* 'bow' in that case. That being said, the use of our contrastive dataset CoMMuTE therefore seems necessary to evaluate how well a MMT model exploits visual information in order to produce correct translations instead of relying only on standard metrics

for MT.

Figure 10 illustrates how VGAMT can translate ambiguous words correctly by using images, while mBART + MT (our strong text-only baseline) cannot. In both cases, the baseline outputs French *noeud papillon* 'bow tie', while VGAMT produces the correct translations of *bow*. Figures 11a to 11f show the same effect for En→{Cs,De} translations. Even if VGAMT does not literally translate the ambiguous word as exemplified by Figure 11b, it produces a translation with the expected meaning based on the image; the text-only models were not able to do so.



Source : We'll have to get rid of that **mole**.

1 Ref : Bude třeba odstranit toto **znaménko**.
VGAMT : Musíme se zbavit toho **znaménka**.
Text-only : Musíme se zbavit toho **znamínka**.

2 Ref : Bude třeba zbavit se tohoto **krtka**.
VGAMT : Musíme se zbavit toho **krtka**.
Text-only : Musíme se zbavit toho **znamínka**.

(a) English word *mole* correctly translated in both cases (*znaménko* 'skin blemish' and *krtka* 'burrowing mammal').

Source : I don't think we should use a **brush**.

1 Ref : Ich glaube nicht, dass wir eine **Bürste** benutzen sollten.
VGAMT : Ich glaube nicht, dass wir eine **Bürste** benutzen sollten.
Text-only : Ich denke nicht, dass wir einen **Pinsel** benutzen sollten.

2 Ref : Ich glaube nicht, dass wir einen **Pinsel** benutzen sollten.
VGAMT : Ich glaube nicht, dass wir einen **Pinsel** benutzen sollten.
Text-only : Ich denke nicht, dass wir einen **Pinsel** benutzen sollten.

(b) English word *brush* correctly translated in both cases (*Bürste* 'cleaning tool' and *Pinsel* 'object used for painting').

Source : They checked the **seal**.

1 Ref : Zkontrolovali **pečeť**.
VGAMT : Oni zkontrolovali **pečeť**.
Text-only : Oni zkontrolovali tu **pečeť**.

2 Ref : Zkontrolovali **tuleně**.
VGAMT : Oni toho **tuleně** prověřili.
Text-only : Oni zkontrolovali tu **pečeť**.

(c) English word *seal* correctly translated in both cases (*pečet'* 'official stamp' and *tuleně* 'sea mammal').

Source : First we should fill up the **boot**.

1 Ref : Zuerst müssen Sie den **Kofferraum** füllen.
VGAMT : Zuerst sollten wir den **Kofferraum** auffüllen.
Text-only : Zuerst sollten wir den **Stiefel** auffüllen.

2 Ref : Zuerst müssen Sie den **Stiefel** füllen.
VGAMT : Zuerst sollten wir den **Stiefel** auffüllen.
Text-only : Zuerst sollten wir den **Stiefel** auffüllen.

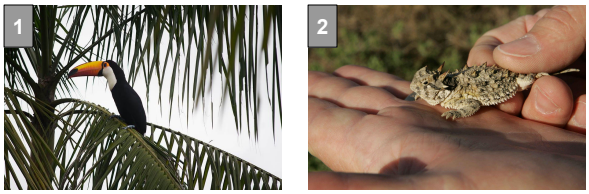
(d) English word *boot* correctly translated in both cases (*Koffer-raum* 'car trunk' and *Stiefel* 'footwear').

Source : Put down your **arms**.

1 Ref : Skloňte **zbraně**.
VGAMT : Položte své **zbraně**.
Text-only : Dej ty **ruce** dolů.

2 Ref : **Připažte.**
VGAMT : Dej **ruce** dolů.
Text-only : Dej ty **ruce** dolů.

(e) English word *arms* correctly translated in both cases (*zbraně* 'weapon' and *ruce* 'parts of the human body').



Source : An animal sitting in a **palm**.

1 Ref : Ein Tier, das auf einer **Palme** sitzt.
VGAMT : Ein Tier sitzt in einer **Palme**.
Text-only : Ein Tier sitzt auf einer **Palme**.

2 Ref : Ein Tier, das in einem **Handteller** sitzt.
VGAMT : Ein Tier sitzt in einer **Handfläche**.
Text-only : Ein Tier sitzt auf einer **Palme**.

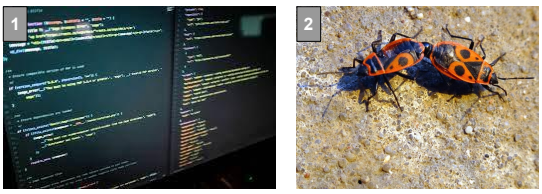
(f) English word *palm* correctly translated in both cases (*Palme* 'tree' and *Handfläche* 'anterior aspect of the hand').

Figure 11: MT examples for different English→Czech and English→German examples from CoMMuTE. For each one, VGAMT is able to exploit visual information to translate English ambiguous words (underlined and in bold) correctly in all cases.

There are some **chips** in a bowl.

1	Il y a des **frites** dans un bol.	**1.192** ✓
	Il y a des **jetons** dans un bol.	1.731
2	Il y a des **jetons** dans un bol.	**1.133** ✓
	Il y a des **frites** dans un bol.	1.547

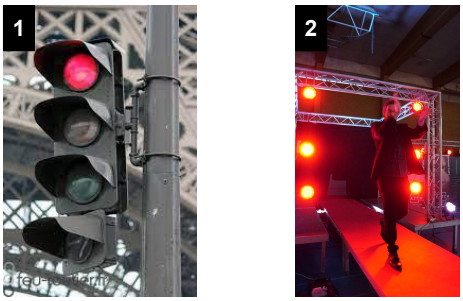
(a) The English word *chips* refers to 'french fries' or 'poker chips'.



There are still a couple of **bugs** left.

1	Il reste encore quelques **bugs**.	**1.217** ✓
	Il reste encore quelques **insectes**.	2.124
2	Il reste encore quelques **insectes**.	**1.331** ✓
	Il reste encore quelques **bugs**.	1.502

(b) The English word *bugs* refers to 'a problem in a computer program' or 'a small insect'.



Can you not see the **red light**?

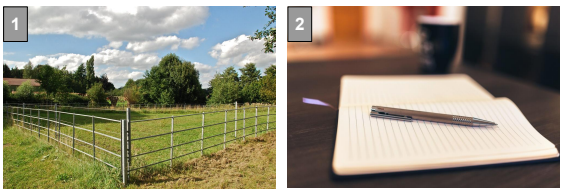
1	Tu ne vois pas le **feu rouge** ?	**1.290** ✓
	Tu ne vois pas la **lumière rouge** ?	1.697
2	Tu ne vois pas la **lumière rouge** ?	**1.324** ✓
	Tu ne vois pas le **feu rouge** ?	1.325

(c) The English phrase *red light* refers to 'a traffic signal that instructs moving vehicles to stop' or 'light that is red'.

The **beam** seems to be moving.

1	Le **rayon** a l'air de bouger.	1.847
	La **poutre** a l'air de bouger.	**1.407** ✗
2	La **poutre** a l'air de bouger.	**1.379** ✓
	Le **rayon** a l'air de bouger.	2.616

(d) The English word *beam* refers to 'a ray of light' or 'a piece of timber or metal used to support the roof'.



Over there by the **pen**.

1	Dort neben dem **Pferch**.	**3.877** ✓
	Dort neben dem **Stift**.	12.635
2	Dort neben dem **Stift**.	**4.772** ✓
	Dort neben dem **Pferch**.	8.258

(e) The English word *pen* refers to 'an instrument for writing or drawing with ink' or 'an area of land surrounded by a fence'.

The **tube** looks dirty!

1	Die **U-Bahn** sieht dreckig aus!	**1.918** ✓
	Das **Röhrchen** sieht dreckig aus!	2.549
2	Das **Röhrchen** sieht dreckig aus!	**2.435** ✓
	Die **U-Bahn** sieht dreckig aus!	4.619

(f) The English word *tube* refers to 'a long cylinder for holding liquids or gases.' or 'a railway system in some cities'.

Figure 12: Perplexity scores from VGAMT on different examples from CoMMuTE. It is possible to produce at least two different French translations from each source sentence in English, the correct translation therefore depends on the input image. For each sub-example, the correct (resp. incorrect) translation is the top (resp. bottom) one. The ambiguous parts of the sentences are highlighted in bold.