Towards Zero-Shot Multimodal Machine Translation

Matthieu Futeral[1,2] **Cordelia Schmid**[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

Current multimodal machine translation (MMT) systems rely on fully supervised data (i.e sentences with their translations and accompanying images), which is costly to collect and prevents the extension of MMT to language pairs with no such data. We propose a method to bypass the need for fully supervised data to train MMT systems, using multimodal English data only. Our method (ZeroMMT) consists in adapting a strong text-only machine translation (MT) model by training it jointly on two objectives: visually conditioned masked language modelling and the Kullback-Leibler divergence between the original MT and new MMT outputs. We evaluate on standard MMT benchmarks and on CoMMuTE, a contrastive test set designed to evaluate how well models use images to disambiguate translations. ZeroMMT obtains disambiguation results close to state-of-the-art MMT models trained on fully supervised examples. To prove that ZeroMMT generalizes to languages with no fully supervised training data, we extend CoMMuTE to three new languages: Arabic, Russian and Chinese. We also show that we can control the trade-off between disambiguation capabilities and translation fidelity at inference time using classifier-free guidance and without any additional data. Our code, data and trained models are publicly accessible.[1,2]

1 Introduction

Multimodal machine translation (MMT) refers to the use of additional modalities, such as images or videos, in machine translation (MT) systems. The main purpose is to provide an additional signal in the case of ambiguity in the text to be translated (i.e the text alone does not provide enough information). Most current MMT models are trained solely on the Multi30K (M30K) dataset (Elliott et al., 2016, 2017; Barrault et al., 2018), a multilingual and multimodal corpus composed of 30K images, their English captions and translations in French, German and Czech. There have been recent breakthroughs in MMT thanks to the use of pretrained text-only MT systems and monolingual captioning data in order to adapt MT systems to MMT (Futeral et al., 2023; Gupta et al., 2023; Vijayan et al., 2024). Good results have been shown using this strategy on CoMMuTE (Futeral et al., 2023), a benchmark designed to evaluate MMT models on their use of images to disambiguate between contrastive translations, and these results were significantly better than MMT systems trained on M30K only (Yin et al., 2020; Yao and Wan, 2020; Liu et al., 2021; Wu et al., 2021; Li et al., 2022b). However, these models still rely on the multilingual *and* multimodal M30K corpus during training to ensure good translation performance. This presents a core limitation: collecting translations of captions is costly,[3] restricting MMT's extension to new languages. Zero-shot transfer between languages has been tested to bypass the problem (Hirasawa et al., 2023), but this results in the poor exploitation of the visual modality to disambiguate ambiguous texts.

In this work, we address this limitation by proposing a method requiring only monolingual multimodal text data (i.e. English text-image pairs), removing the need for fully supervised data, i.e. parallel and multimodal data such as M30K. We start from a strong pretrained MT system and use it to translate multimodal English data into the target languages of interest. We then adapt the pretrained MT system to images using two objectives: (1) visually conditioned masked language modelling (VMLM) (Li et al., 2019; Lu et al., 2019) on multimodal English data to force the model to use image

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

[3]Authors of M30K stated they spent €23,000 on the translation of the 30,000 English captions into German.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 761–778

information and (2) a KL penalty on the translated multimodal data to maintain translation capabilities. We test our method on six languages directions: English to French, Czech, German, Arabic, Russian and Chinese, extending the CoMMuTE dataset to cover the three additional languages. Our method, called ZeroMMT, obtains CoMMuTE scores close to the supervised state of the art, while there is only a small drop in BLEU and COMET scores compared to the underlying text-only MT system on standard MMT benchmarks composed mainly of unambiguous examples (i.e. where images are not useful for correct translation). We further show that we can control the trade-off between disambiguation and general translation performance at inference time with classifier-free guidance.

2 Related Work

Training MMT systems Research in MMT originally focused on which visual features to use (Li et al., 2022a) and how to integrate them into sequence-to-sequence models (Sutskever et al., 2014) trained from scratch on the widely used M30K benchmark (Libovický et al., 2016; Calixto et al., 2016; Elliott and Kádár, 2017; Calixto and Liu, 2017; Yin et al., 2020; Liu et al., 2021; Li et al., 2022b). These MMT systems typically show improvements of around 1-2 BLEU points on standard MMT benchmarks in comparison to text-only baselines trained from scratch, which is not significant enough to state that MMT systems are better than their text-only counterparts (Mathur et al., 2020). Wu et al. (2021) observed that while they obtained +1 BLEU on average on M30K test sets with the use of images, they got the same improvements with randomly initialized visual features, most likely due to regularization, i.e. the images were in reality not being exploited effectively. On top of that, being trained from scratch on fully supervised MMT data only, these models lag far behind state-of-the-art MT systems (Costa-jussà et al., 2022) trained on large amounts of parallel text.

Futeral et al. (2023) show that M30K contains few ambiguous examples requiring visual context, and that models can get good results on the benchmark while still struggling to exploit images correctly. They introduce VGAMT, an adapted MMT model based on a frozen state-of-the-art MT model. They also show that visually masked language modelling (VMLM) on English captioning data was a key additional objective to force MMT systems to become truly multimodal. Sato et al. (2023) and Bowen et al. (2024) further show that choosing the masked tokens in a smart way instead of randomly slightly boosts results. However, these methods still require fully supervised data to be good at translation; training on VMLM alone results in a collapse in translation capabilities.

A few works have used pseudo-multimodal parallel data by translating English captions into the target language using a pretrained MT system (Li et al., 2021; Caglayan et al., 2021; Vijayan et al., 2024). However, Caglayan et al. (2021) and Vijayan et al. (2024) used them in a pretraining step in a form of distillation of the knowledge of the MT system into the new MMT model before finetuning on M30K. Li et al. (2021) use backtranslation to translate English captions into Turkish to train a Turkish-to-English MMT model for disambiguating gender pronouns from Turkish to English. While effective, their method cannot be applied beyond this particular context because it requires the MT system to output the correct translation, which cannot be assumed to be the case in more general ambiguous contexts (i.e. when text context is not enough to translate the English text correctly).

There have been efforts to train MMT models without using fully supervised data (Su et al., 2019; Huang et al., 2020; Fei et al., 2023). These approaches are however fundamentally different from this work as their goal is to obtain MT models using synthetic text-only parallel data through the use of visual pivoting, not targeting disambiguation capabilities. Hirasawa et al. (2023) proposed a zero-shot method to learn MMT by training on the little fully supervised data available aiming for zero-shot cross-lingual transfer. As the amount of fully supervised data for a single language is small ($\leq$ 30K text-image pairs), and few languages are covered ($\leq$ 8), this method results in poor exploitation of the image to learn disambiguation capabilities.

Evaluating MMT systems The test sets typically used to evaluate MMT systems are the test subsets of M30K (Elliott et al., 2016, 2017; Barrault et al., 2018). However, some of the translations were produced without access to the images and they have also been found to contain only a few ambiguous examples where visual context is necessary (Futeral et al., 2023). They are therefore not best adapted to evaluating MMT systems. Elliott (2018) and Caglayan et al. (2019) proposed to use an adversarial evaluation method and a probing method

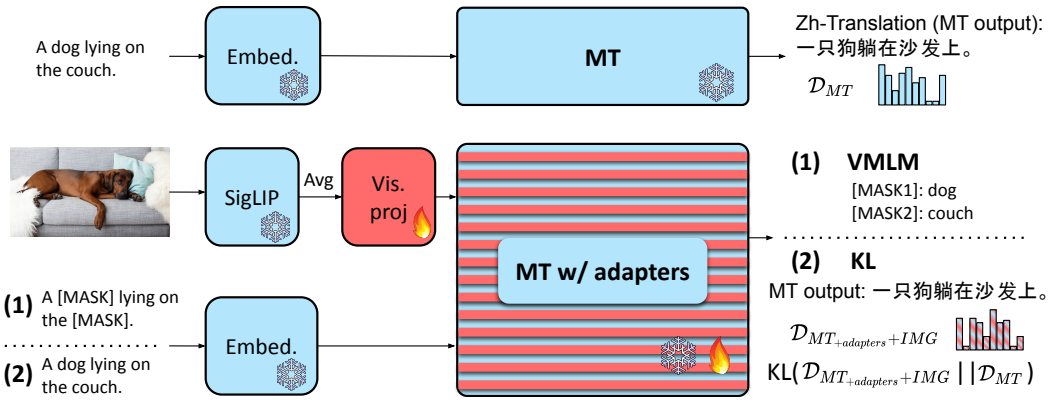


Figure 1: Overview of our approach. We train on two objectives: Visually conditioned masked language modelling (VMLM) and Kullback-Leibler (KL) divergence. All weights are frozen during training except the visual projector and the adapters in the MT model.

based on masked text inputs to assess the utility of images in translation. However, this is not a good proxy for evaluating an MMT model's capacity to disambiguate translations as it relies on masked inputs and so says little about models' capacities to disambiguate when given unmasked text inputs. Lala and Specia (2018), Li et al. (2021) and Zhu et al. (2023) released evaluation datasets composed of sentences in English with ambiguous words accompanied with disambiguating images. However, Li et al. (2021) only target the ambiguity of gender pronouns, and all these datasets are prone to distributional bias, which is difficult to measure and is such that text-only MT systems can perform very well on them (i.e. images are in fact often not necessary for correct translation). Traditional MT metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Rei et al., 2020) are also unable to catch how well MMT systems use images as they do not specifically target translations where images would be required to translate correctly. Tackling these issues, Futeral et al. (2023) introduced CoMMuTE, a contrastive evaluation dataset, composed of English sentences to be translated, built around ambiguous words, each sentence accompanied by two translations with two images, each of which disambiguates the English sentence. MMT models are evaluated on their capacity to give a lower perplexity to the correct translation than the incorrect one, given the source sentence and an image. As perplexity is used to evaluate MMT models, it is a direct proxy of MMT models' capacity to disambiguate English sentences. Furthermore, text-only MT systems can only perform as well as random (50%), as they do not have access to the images.

3 Extending the CoMMuTE benchmark

	En	Ar	Ru	Zh
#unique sents.	155	310	310	310
#tokens	1,384	2,958	3,105	2,832
#unique toks.	559	870	1,002	762

Table 1: Statistics of the extension of CoMMuTE.

Currently available for English-to-{French,German,Czech}, we extend the CoMMuTE benchmark to three new target languages: Arabic, Chinese and Russian, using professional translators. We also release a small validation set of 30 English ambiguous words (non-overlapping with test set examples) with two French translations, each with its own image, to be used for model selection during training. Table 1 shows statistics of our extension of CoMMuTE.[4]

4 Our Approach

Our goal is to train an MMT model capable of effectively using images to disambiguate ambiguous translations (i.e. where an image is necessary to translate correctly, which is MMT's main purpose) while keeping the general MT capacity of the underlying MT model, without using fully supervised data (i.e. in a zero-shot way). This allows us to extend MMT to more language pairs, currently not possible without collecting fully supervised data.

As shown in Figure 1, we start from a strong pretrained NLLB (Costa-jussà et al., 2022) MT model and use it to translate English captions. Similarly to Futeral et al. (2023), we turn it into an

[4]Tokenisation using NLLB (Costa-jussà et al., 2022).

MMT model by adding lightweight trainable modules (visual projectors and adapters), keeping original weights frozen during training. We use visual embeddings from SigLIP (Zhai et al., 2023) and concatenate them to the sequences of text embeddings in the NLLB encoder. Our approach, ZeroMMT, is based on two objectives: (1) force the model to use images when translating by using visually-conditioned masked language modeling (VMLM) and (2) maintain the performance of the original MT system without any fully supervised data using the Kullback-Leibler (KL) divergence between the MMT system's output and the original MT system's output distributions using the previously automatically translated data. While (1) has already been proved successful for learning visual disambiguation capabilities (Futeral et al., 2023), we further show (Section 6) that (2) is key for retaining a strong translation capacity in an MMT setting, enabling zero-shot training.

In more detail, let $x_{1,\dots,n}$ denote the sequence of tokens of the English sentence, i the image embedding, $y_{1,\dots,m}$ the translated sequence of tokens, f_θ the original MT system and $f_{\theta,\beta}$ the MMT system built on top of the text-only MT model with additional light-weight modules β, both outputting probability distributions over tokens. We formally define the losses as follows:

$$\mathscr{L}_{VMLM} = \sum_j y_j \log \left(f_{\theta,\beta}(y_j; y_{<j}, x_{\setminus \mathscr{M}}, i) \right) \quad (1)$$

$$\mathscr{L}_{KL} = \sum_j f_\theta(y_j; y_{<j}, x) \log \frac{f_\theta(y_j; y_{<j}, x)}{f_{\theta,\beta}(y_j; y_{<j}, x, i)} \quad (2)$$

where $\mathscr{M}$ is the set of masked input indices. The final loss is a weighted combination of (1) and (2), and we choose the λ value based on results on validation sets as described in Section 5.2:

$$\mathscr{L} = \mathscr{L}_{VMLM} + \lambda \mathscr{L}_{KL}$$

5 Experiments

5.1 Data

We trained our models on the Conceptual Captions dataset[5] (Sharma et al., 2018). We translated Conceptual Captions into French, German, Czech, Chinese, Russian and Arabic using NLLB (Costa-jussà et al., 2022) (of size 600M, 1.3B or 3.3B depending on the experiment) using a beam of size 4 for the 600M model and 2 for the largest ones.

We evaluate our models on the M30K test sets (Elliott et al., 2016, 2017; Barrault et al., 2018) for English-to-{German,French,Czech}, the EMMT test set (Zhu et al., 2023) for English-to-Chinese, comprising 500 English product titles from e-commercial websites translated into Chinese, and the VATEX test set (Wang et al., 2019) for English-to-Chinese, composed of 10-second videos[6] with English captions translated into Chinese. We use these test sets to make sure general translation quality is not harmed when introducing additional visual inputs in unambiguous cases (as described previously, they cannot be used in practice to evaluate MMT models' ability to use images correctly). Finally we evaluate on CoMMuTE for English-to-{German,French,Czech,Chinese,Russian,Arabic}, used to test how well the MMT models exploit visual context for disambiguation.

5.2 Implementation details

Modelling We trained three different versions of ZeroMMT depending on the size of the underlying NLLB model[7] (Costa-jussà et al., 2022) (600M, 1.3B and 3.3B). For SigLIP (Zhai et al., 2023), we use `ViT-B-16-SigLIP-384` trained on WebLI (Chen et al., 2023) from the `timm` library (Wightman, 2019). Following VGAMT (Futeral et al., 2023), we used bottleneck adapters (Houlsby et al., 2019) as implemented in the Adapters Python library (Poth et al., 2023) with a factor reduction of 8 and ReLU activation (Agarap, 2018) for each layer. The visual projector is a 1-layer neural network followed by ReLU activation projecting SigLIP (Zhai et al., 2023) embeddings towards the hidden dimension of NLLB. The image representation is then concatenated to the sequence of text embeddings. The cross-attention mechanism in the decoder of the model can only attend to the positions of text embeddings. Similarly to VGAMT, we randomly mask 25% of the input tokens for VMLM.

Training We train our models with a batch size of 32, the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and learning rate of 10^{-4}. We use $\lambda = 0.1$ to balance the two training losses. All hyperparameters were selected based on the combination of the CoMMuTE validation set (see Section 3) and the English–French valida-

[5] At the time of writing, we were able to collect 2,831,746 out of the 3,300,000 images.

[6] We take 5 frames per second, compute SIGLIP features and average them to obtain the visual input.

[7] As implemented in `Transformers` (Wolf et al., 2020).

	Ar	Cs	De	Fr	Ru	Zh
Text-only MT baselines	*50.0*	*50.0*	*50.0*	*50.0*	*50.0*	*50.0*
NLLB-SIGLIP *topline*	82.6	76.0	83.7	75.0	75.8	88.1
MMT – *fully supervised*						
Gated Fusion *bilingual*	-	51.0 ±1.9	49.7 ±0.6	50.0 ±0.8	-	-
VTLM + MMT *bilingual*	-	52.0 ±0.7	50.2 ±0.3	51.4 ±0.9	-	-
VGAMT *full bilingual*	-	55.6 ±0.8	**59.0** ±0.5	**67.1** ±0.7	-	-
VGAMT *SIGLIP-only multi.*	-	**57.5** ±1.2	57.1 ±0.4	61.3 ±1.1	-	-
MMT – *cross-lingual zero-shot*						
M2KT-VPN *bilingual*	-	50.1 ±0.6	50.3 ±0.5	50.9 ±0.8	-	-
MMT – *zero-shot*						
Multilingual OpenFlamingo	**61.3**	59.1	**63.7**	**68.5**	**67.4**	**66.5**
ZeroMMT-600M (*ours*) *multi.*	56.1±0.8	55.5±0.5	55.7±0.3	58.7±0.4	57.2±1.2	58.2±1.1
ZeroMMT-1.3B (*ours*) *multi.*	57.3±0.2	59.4±0.5	57.4±0.4	62.2±0.5	60.6±0.5	60.1±0.8
ZeroMMT-3.3B (*ours*) *multi.*	58.9±0.5	**61.7**±0.3	60.8±0.8	65.0±0.7	62.9±0.3	60.1±0.7

Table 2: Results on CoMMuTE, averaged over 3 runs (± standard error). The best scores for each category are in **bold** and the second best are underlined.

	Fr		De		Cs		Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Text-only MT baselines								
NLLB-600M *distilled*	49.17 ±0.78	85.18 ±0.67	33.04 ±3.44	81.98 ±2.16	26.58 ±0.19	85.02 ±0.42	16.07 ±1.35	57.81 ±4.21
NLLB-1.3B	51.90 ±0.79	86.28 ±0.77	35.39 ±2.83	83.49 ±2.15	30.77 ±0.46	87.48 ±0.29	18.22 ±0.20	60.02 ±3.51
NLLB-3.3B	53.73 ±0.57	86.98 ±0.88	37.26 ±2.10	84.76 ±1.76	33.37 ±0.27	88.70 ±0.37	20.55 ±0.46	61.27 ±3.50
MMT – *fully supervised*								
Gated Fusion *bilingual*	49.79 ±7.46	80.62 ±3.01	31.57 ±5.24	72.89 ±3.15	28.30 ±2.52	79.24 ±2.41	-	-
VTLM + MMT *bilingual*	55.27 ±6.00	83.45 ±1.98	35.94 ±3.44	79.10 ±2.35	32.63 ±2.26	82.40 ±1.77	-	-
VGAMT *full bilingual*	59.97 ±6.66	88.29 ±1.83	39.10 ±3.14	85.72 ±1.73	35.89 ±1.70	89.50 ±1.08	-	-
VGAMT *SIGLIP-only multi.*	58.39 ±5.67	87.27 ±1.74	37.36 ±3.51	83.85 ±2.04	34.88 ±1.77	87.45 ±1.19	-	-
MMT – *cross-lingual zero-shot*								
M2KT-VPN *bilingual*	51.58 ±6.72	80.19 ±3.77	29.27 ±5.77	71.63 ±2.77	28.02 ±2.31	78.63 ±2.77	-	-
MMT – *zero-shot*								
Multilingual OpenFlamingo	35.08 ±0.76	82.66 ±1.38	24.92 ±2.89	79.93 ±2.44	3.27 ±0.04	70.73 ±0.55	8.60 ±5.86	53.38 ±10.24
ZeroMMT-600M (*ours*) *multi.*	49.00 ±1.07	84.82 ±0.79	32.79 ±2.97	81.13 ±2.48	25.24 ±0.62	83.79 ±0.55	15.74 ±1.62	57.10 ±4.72
ZeroMMT-1.3B (*ours*) *multi.*	52.06 ±1.15	86.15 ±0.84	35.18 ±2.58	83.35 ±1.90	30.14 ±0.48	86.94 ±0.33	17.11 ±0.71	59.17 ±4.34
ZeroMMT-3.3B (*ours*) *multi.*	53.34 ±0.50	86.69 ±0.94	37.08 ±2.49	84.41 ±1.77	33.03 ±0.34	88.37 ±0.32	19.43 ±0.64	60.61 ±4.28

Table 3: Aggregated generation results for En→X. Fr and De results are averaged over Test2016, Test2017 from M30K and AmbiguousCOCO. Cs results are averaged over M30K Test2016 and Test2018. Zh results are averaged over EMMT and VATEX test sets.

tion dataset of M30K, each score weighted equally. All our models are multilingual if not otherwise specified. We run each experiment three times with three different seeds and report average scores and standard error. It took 15 hours on one NVIDIA V100 for the 600M model and 20 hours on one NVIDIA A100 for the largest models.

Evaluation We evaluate MMT generation with BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). For BLEU, we use the Sacrebleu implementation (Post, 2018) with *13a* tokenization for French, German and Czech and *zh* tokenization for Chinese. For COMET, we use the `wmt22-comet-da` (Rei et al., 2022) model from the XLM-R backbone (Conneau et al., 2020). The translations were obtained with beam search decoding of size 4. Following (Futeral et al., 2023), we calculate disambiguation accuracy using CoMMuTE: given an English sentence and an associated image, we compute the perplexities of each contrastive translation, giving a score of 1 if the perplexity of the correct translation is lower than the perplexity of the contrastive one and 0 otherwise.

5.3 Results

Baselines and comparative models We compare our approach to several others. Firstly, we compare to the text-only MT systems on which the ZeroMMT models are based, NLLB-600M *distilled*, NLLB-1.3B and NLLB-3.3B. We also compare against well-known fully supervised MMT systems: Gated Fusion (Wu et al., 2021), a tiny 3M-parameter Transformer (Vaswani et al., 2017) trained from scratch on M30K; VTLM (+ MMT) (Caglayan et al., 2021), a 44M-parameter MMT

	Ar	Cs	De	Fr	Ru	Zh
NLLB-600M	79.06	**83.31**	80.84	80.17	79.94	75.41
ZeroMMT-600M (*ours*)	**79.43**	82.62	**81.11**	**81.36**	**80.67**	**75.90**
NLLB-1.3B	80.59	84.12	81.28	80.14	81.72	75.30
ZeroMMT-1.3B	**81.02**	**85.41**	**83.46**	**82.35**	**83.45**	**77.06**
NLLB-3.3B	79.90	84.97	81.20	81.16	82.26	76.69
ZeroMMT-3.3B	**80.64**	**86.69**	**83.54**	**83.40**	**83.87**	**77.85**

Table 4: COMET scores on CoMMuTE (as a translation set). The best result for each model size is in **bold**.

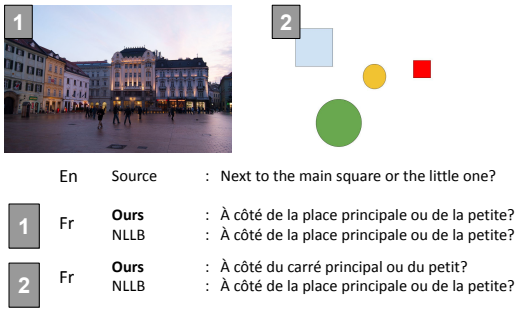
system first pretrained on the translation language modelling (TLM) objective (Lample and Conneau, 2019) with an additional image as input on translated captioning data (using the same translated data and tokenizer as ZeroMMT-600M) and then MMT-finetuned on M30K; and VGAMT (Futeral et al., 2023), a 630M-parameter MMT system (of which 13M trainable), which is an MT-fine-tuned mBART (Liu et al., 2020) transformed into an MMT system through the addition of lightweight adapters trained jointly on the MMT and VMLM objectives. VGAMT originally uses multiple types of visual input and is bilingual. Therefore, to have a comparable setup we retrain a VGAMT-like model with NLLB-600M *distilled* as the underlying MT model, with SIGLIP features only and in a multilingual setting. Finally, we compare to Multilingual OpenFlamingo (Futeral et al., 2024), a 3B multilingual multimodal language model pretrained on a large number of text-image pairs and interleaved documents which allows for zero-shot MMT in a way that is comparable with our model and M2KT-VPN (Hirasawa et al., 2023), a cross-lingual zero-shot MMT model based on a tiny Transformer.

We compute an approximate upperbound on CoMMuTE for models trained with SIGLIP features by evaluating on NLLB-SIGLIP (Visheratin, 2023). For each CoMMuTE instance, we compute the cosine similarity between the translation and (i) its associated image and (ii) the other image. If the cosine similarity of (i) is higher than (ii), it is considered a correct prediction.

Quantitative results Tables 2 and 3 show the results on CoMMuTE and the aggregated results on generation benchmarks not composed of ambiguous examples (i.e. images are not required) respectively. For full results, see Appendix A. Compared to the text-only NLLB-600M distilled model, our approach results in only a small drop in performance on generation benchmarks (-0.52 BLEU and -0.79 COMET on average), where images are not required to translate the sentence correctly, despite not using the M30K training data or any fully supervised data. For the disambiguation task, Multilingual OpenFlamingo obtains the strongest CoMMuTE scores but it fails in generation as it was not specifically trained to translate. Our approach is significantly better than the random baseline (>55% for all languages for the smallest model, >61% on average for the largest model), showing that it is able to exploit images for disambiguation; results are close to VGAMT scores (for similar model sizes) for Czech despite not having been trained on fully supervised data. Table 4 shows additional results on CoMMuTE but used as a traditional MMT generation benchmark. We obtain higher COMET scores than the text-only baseline NLLB on all languages for all model sizes except Czech for the smallest model. These results show that our approach is able to improve translation performance by exploiting images for disambiguation in cases of ambiguous examples without using any fully supervised data during training. We shall see in Section 7 how image exploitation can be controlled at inference time and how our approach can be made to outperform Multilingual OpenFlamingo in image exploitation (as measured on CoMMuTe).

Qualitative results We analysed some translations of our ZeroMMT-600M model and compared them with those of the text-only distilled NLLB-600M model. Our model is able to exploit the image to slightly change the translation towards the correct meaning, as shown in Figure 2a, where ambiguous parts of the translations change when the image is provided. In Figure 2b, the translation is also improved; *bass* is translated as 鱼 'fish' rather than 低音 'bass (low tone)'. We also notice few variations in the other areas of the translation with respect to the NLLB translation, which means that our model correctly identifies the part to change. More examples can be found in Appendix B.

Human evaluation We conduct human evaluation between ZeroMMT-3.3B and NLLB-3.3B outputs to further confirm these results. We set up A/B testing where annotators were asked to assess which of the translations was better given the source translation and the accompanying image. We randomly sampled 100 examples from CoMMuTE to assess which model is better in cases of ambiguity in the source sentence and 100 examples from M30K test sets (equally represented) in cases where images do not provide additional informa-



(a) English–French example.

(b) English–Chinese example.

Figure 2: Translations of CoMMuTE by our approach, ZeroMMT-600M, and the NLLB distilled MT model.

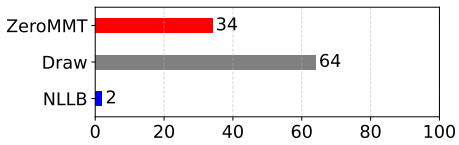


Figure 3: French human evaluation on 100 examples from CoMMuTE. 'Draw' means both translations are exactly the same or considered of same quality.

tion for translation. Figure 3 shows that ZeroMMT-3.3B is considered to be better than NLLB-3.3B by a large margin in cases of ambiguity. In cases where the image adds no additional information (i.e. the sentence to translate is non-ambiguous), Figure 4 shows that translations are considered the same in most cases (80%), and in the remaining cases NLLB-3.3B is considered to be slightly better than ZeroMMT-3.3B (11 vs. 9). The results of a similar analysis for Arabic and Chinese is given in Figures 5 to 7. By manually looking at some examples, we noticed that the few unambiguous cases where NLLB-3.3B is considered superior to ZeroMMT-3.3B are due to hallucinations (since we add additional information in cases it is not necessary, this can occasionally occur).

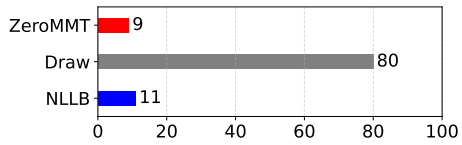


Figure 4: French human evaluation on 100 examples from M30K test sets. 'Draw' means both translations are exactly the same or considered of the same quality.

6 Ablation study

	Translation sets BLEU	Translation sets COMET	CoMMuTE accuracy
ZeroMMT-600M	<u>32.73</u> ±12.33	<u>77.95</u> ±10.82	<u>56.9</u> ±1.4
w/o VMLM	**33.12** ±12.01	**78.49** ±10.69	50.3 ±0.4
w/o KL	14.10 ±10.70	65.88 ±11.72	**58.9** ±1.8
+ MMT w/o KL	32.09 ±12.62	77.50 ±10.80	55.5 ±1.3

Table 5: Ablation study. Aggregated scores over benchmarks and languages. The best results are in **bold** and second best are <u>underlined</u>.

We conduct an ablation study on our ZeroMMT-600M model to analyse the impact of our two objectives. We first train a model without the VMLM objective, then a model without the KL penalty. We also test the replacement of the KL penalty with a standard auto-regressive MMT translation loss with the translated data as the ground truth, and finally we vary the KL penalty coefficient to observe the evolution of COMET and CoMMuTE scores. Additional ablation study on the choice of visual feature can be found in Appendix A.4.

KL penalty only (i.e. without VMLM) Table 5 shows that with the KL penalty only, the model cannot exploit visual information for translation. This is because there is no need to use the input image and the model learns to ignore it. The aggregated CoMMuTE score is close to random guessing.

VMLM only (i.e. without KL) Table 5 also shows that, while the VMLM objective allows the model to obtain good scores on CoMMuTE (it is able to exploit visual information), the scores on generation benchmarks collapse as expected, with -19 BLEU points and -12 COMET points in comparison to the full approach.

KL penalty vs. MMT objective Finally, we replace the KL penalty with the standard MMT objective (i.e. *+MMT w/O KL* in Table 5) as the objective to maintain translation quality. We observe a drop of 0.64 BLEU points and 0.45 COMET

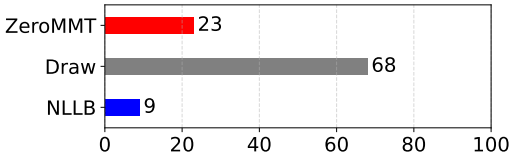


Figure 5: Chinese human evaluation on 100 examples from CoMMuTE. 'Draw' means both translations are exactly the same or considered of same quality.

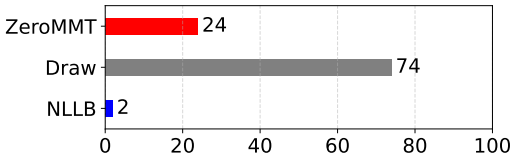


Figure 7: Arabic human evaluation on 100 examples from CoMMuTE. 'Draw' means both translations are exactly the same or considered of the same quality.

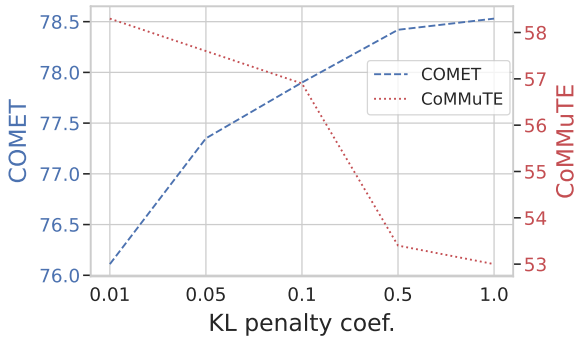


Figure 8: Evolution of aggregated COMET and CoMMuTE scores when changing the KL penalty coefficient.

points on average in comparison with the use of the KL penalty. It additionally results in an average drop of 1.4 points on CoMMuTE.

Varying the trade-off between objectives In Figure 8 we show the variation of COMET and CoMMuTE when testing our approach with different λ coefficients for the KL penalty. We notice that when λ is too high, it results in a large average drop of performance on CoMMuTE.

7 Controlling the disambiguation level

We show that our method allows us to obtain an MMT system with a good trade-off between strong translation quality on unambiguous examples (i.e. where images are not necessary to translate correctly) and the capacity to exploit visual context for disambiguation. However, some applications could require stronger disambiguation capabilities and be less reliant on translation fidelity on unambiguous cases or vice versa. Instead of retraining a model to control the trade-off between the two objectives, we instead propose to use classifier-free guidance (CFG) (Ho and Salimans, 2021; Sanchez et al., 2023) to control this trade-off at inference time. We define CFG in the context of MMT as follows:

$$\widehat{f}_{\theta,\beta}(y_j; y_{<j}, x, i) = f_\theta(y_j; y_{<j}, x) + \gamma\big(f_{\theta,\beta}(y_j; y_{<j}, x, i) - f_\theta(y_j; y_{<j}, x)\big) \quad (3)$$

where f_θ is the text-only MT system, $f_{\theta,\beta}$ the adapted MMT system, x and y the source and generated sentence, i the visual input, j the token index and γ the CFG value controlling guidance.

We analyse the evolution of BLEU and COMET scores on standard generation benchmarks (where text context is enough to translate correctly), and CoMMuTE scores when varying the γ parameter. Table 6 shows that ZeroMMT-3.3B can achieve a boost in CoMMuTE accuracy of up to 4.2 points for $\gamma = 1.5$, while facing only a moderate drop of BLEU and COMET scores on unambiguous generation benchmarks (which do not require images as additional context in theory). Higher γ values result in stronger disambiguation capabilities, as shown by CoMMuTE, but this comes at the ex-

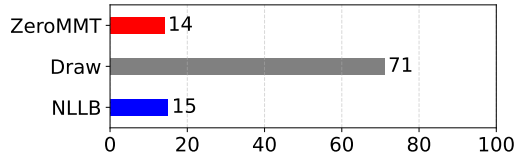


Figure 6: Chinese human evaluation on 100 examples from VATEX. 'Draw' means both translations are exactly the same or considered of same quality.

γ	BLEU	COMET	CoMMuTE
1.0	**37.62** ±12.11	**81.12** ±10.59	61.6 ±2.1
1.25	<u>37.30</u> ±12.07	<u>80.98</u> ±10.56	64.2 ±2.5
1.5	36.84 ±11.89	80.76 ±10.53	65.8 ±2.7
2.0	35.02 ±11.07	79.92 ±10.42	68.5 ±2.9
2.5	32.15 ±10.14	78.42 ±10.16	<u>70.3</u> ±2.6
3.0	28.75 ±9.08	76.25 ±9.69	**71.7** ±2.5
MOF	*20.37 ±12.92*	*73.60 ±11.98*	*64.4 ±3.4*

Table 6: Evolution of BLEU, COMET and CoMMuTE scores of our CFG-enabled ZeroMMT-3.3B model (aggregated over benchmarks and languages) compared to CFG-free ZeroMMT (i.e. γ = 1.0). The **best** and <u>second best</u> results for each model size are shown typographically. "MOF" shows scores for Multilingual OpenFlamingo.

pense of a drop in generation quality on the unambiguous benchmarks. CFG can therefore allow us to control the trade-off between disambiguation capability and translation fidelity depending on the application. Importantly, we strongly outperform Multilingual OpenFlamingo on all metrics for different CFG values and we can obtain CoMMuTE scores up to 71.7 on average for $\gamma = 3.0$. Results for ZeroMMT-600M and 1.3B can be found in Table 11 in Appendix A.2.

8 Conclusion

We present ZeroMMT, a novel zero-shot MMT approach bypassing the need for parallel multimodal data. ZeroMMT shows good disambiguation capabilities (it is able to effectively exploit images) while maintaining good translation results, with only a very small drop in performance according to standard generation benchmarks where images are not necessary for correct translation. ZeroMMT allows us to extend MMT to new language directions; we show that it performs well on the CoMMuTE test set for Russian and Arabic for which no parallel multimodal training data is available. Moreover, we show that it is possible to control the disambiguation-generation trade-off using classifier-free guidance. It is therefore a step towards having MMT systems that cover a broader set of languages without having to rely on acquiring costly training data.

Limitations

While our approach allows us to exploit images for translation disambiguation as shown by the scores obtained on CoMMuTE, it is still behind the upperbound. Zero-shot disambiguation capabilities also come at the expense of a slight drop in translation quality in cases where text context is enough to translate correctly as shown by BLEU and COMET scores. To fill this gap, a next step, which we leave to future work, would be to detect ambiguity in the source sentence and access the images only in those cases. Indeed, in most cases, images are not necessary to translate the English source sentence correctly. There are therefore areas for improvement even if our zero-shot approach is close to its fully supervised counterparts. It is nevertheless a step towards zero-shot multimodal machine translation and the expansion of MMT to new language pairs.

Ethics Statement

The released extension of CoMMuTE is designed to evaluate disambiguation capabilities of MMT systems and should not be used in any other way. Images were collected under the Creative Commons license and CoMMuTE is distributed under CC-BY-SA-4.0 license. All of our models are also distributed under CC-BY-SA-4.0 license.

References

Abien Fred Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375*.

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.

Braeden Bowen, Vipin Vijayan, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Detecting concrete visual tokens for multimodal machine translation. *arXiv preprint arXiv:2403.03075*.

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, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, Berlin, Germany. 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.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali Rwanda.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

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.

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

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

Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. 2024. mOSCAR: A large-scale multilingual and multimodal document-level corpus. *arXiv preprint arXiv:2406.08707*.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. CLIPTrans: Transferring Visual Knowledge with Pre-trained Models for Multimodal Machine Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Tosho Hirasawa, Emanuele Bugliarello, Desmond Elliott, and Mamoru Komachi. 2023. Visual prediction improves zero-shot cross-modal machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 522–535, Singapore. Association for Computational Linguistics.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

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.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada.

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

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

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. *CoRR*, abs/1908.03557.

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

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 646–654, Berlin, Germany. Association for Computational Linguistics.

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 pre-training 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.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. 2023. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*.

Julia Sato, Helena Caseli, and Lucia Specia. 2023. Choosing what to mask: More informed masking for multimodal machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 244–253, Toronto, Canada. 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.

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10482–10491.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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*, pages 5998–6008.

Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Adding multimodal capabilities to a text-only translation model. *arXiv preprint arXiv:2403.03045*.

Alexander Visheratin. 2023. NLLB-CLIP–train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*.

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

Ross Wightman. 2019. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`.

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.

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.

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.

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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

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

A Detailed results

A.1 Main results

Tables 7 to 10 show full BLEU and COMET scores for all languages and all benchmarks.

A.2 Impact of γ in CFG

Table 11 shows the impact of γ on BLEU, COMET and CoMMuTE scores of our CFG-enabled ZeroMMT 600M and 1.3B models (aggregated over benchmarks and languages) compared to the vanilla, CFG-free ZeroMMT model (i.e. $\gamma = 1.0$). See Section 7 for results with our ZeroMMT-3.3B model.

A.3 Ablation study - Full results

Tables 12 to 15 show the full results of the ablation study for all languages and all benchmarks.

A.4 Ablation study - Choice of visual representation

We additionally train several ZeroMMT-600M models with different types of visual encoder. As shown by Table 16, the type of visual encoder does not have an impact on global translation performances as BLEU and COMET scores do not vary a lot on standard benchmarks between models. However, we notice significant differences on CoMMuTE; the performance of CLIP,[8] SIGLIP[9] and SIGLIP large[10] visual encoders are about 1.5 points higher on average on CoMMuTE in comparison to VIT[11] and ResNet-50.[12] This is probably due to the fact that VIT and ResNet-50 are trained on ImageNet, which limits their capacity to ImageNet classes while CLIP and SIGLIP-like visual encoders are trained on free-form image-text large datasets. However, all scores on CoMMuTE are well above random, therefore validating the method for different types of visual representation.

B Additional examples

Figures 9a to 9f show additional translation examples from CoMMuTE by ZeroMMT (Ours) and the text-only NLLB-600M distilled model.

[8] `ViT-B-32`
[9] `vit_base_patch16_siglip_384`
[10] `vit_so400m_patch14_siglip_384`
[11] `google/vit-base-patch16-224-in21k`
[12] `microsoft/resnet-50`

	Test2016		Test2017		COCO	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Text-only MT baselines						
NLLB-600M *distilled*	48.71	85.18	48.54	85.99	50.28	84.36
NLLB-1.3B	51.68	86.60	51.06	87.01	52.95	85.22
NLLB-3.3B	54.15	87.57	52.92	87.64	54.11	85.73
MMT – *fully supervised*						
Gated Fusion *bilingual*	58.70 ± 0.30	83.60 ± 0.08	50.80 ± 0.70	81.74 ± 0.25	40.40 ± 0.40	76.52 ± 0.25
VTLM + MMT *bilingual*	63.37 ± 0.13	85.29 ± 0.06	55.77 ± 0.17	84.35 ± 0.04	47.69 ± 0.16	80.70 ± 0.20
VGAMT *full bilingual*	67.20 ± 0.10	89.78 ± 0.04	61.60 ± 0.10	89.37 ± 0.04	51.10 ± 0.60	85.78 ± 0.11
VGAMT *SIGLIP-only multi.*	65.04 ± 0.52	88.74 ± 0.04	58.90 ± 0.28	88.23 ± 0.19	51.24 ± 0.73	84.84 ± 0.29
MMT – *cross-lingual zero-shot*						
M2KT-VPN *bilingual*	59.21 ± 0.56	83.95 ± 0.13	52.63 ± 0.63	81.55 ± 0.41	42.90 ± 0.20	75.08 ± 0.51
MMT – *zero-shot*						
Multilingual OpenFlamingo	36.01	83.56	35.10	83.72	34.14	80.71
ZeroMMT-600M (*ours*) *multi.*	48.62 ± 0.38	84.92 ± 0.09	48.10 ± 0.11	85.66 ± 0.16	50.29 ± 0.82	83.78 ± 0.20
ZeroMMT-1.3B (*ours*) *multi.*	51.47 ± 0.11	86.42 ± 0.17	51.10 ± 0.02	87.00 ± 0.17	53.60 ± 0.54	85.03 ± 0.08
ZeroMMT-3.3B (*ours*) *multi.*	52.89 ± 0.36	87.22 ± 0.05	53.29 ± 0.19	87.48 ± 0.13	53.86 ± 0.30	85.38 ± 0.13

Table 7: En→Fr results for Test2016, Test2017 and COCO subsets of M30K, avg. over 3 runs (± standard error).

	Test2016		Test2017		COCO	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Text-only MT baselines						
NLLB-600M *distilled*	37.14	83.79	33.24	83.21	28.73	78.95
NLLB-1.3B	37.91	85.14	36.81	84.86	31.44	80.45
NLLB-3.3B	39.47	86.22	37.86	85.76	34.44	82.28
MMT – *fully supervised*						
Gated Fusion *bilingual*	38.70 ± 0.20	76.32 ± 0.17	29.50 ± 0.20	73.61 ± 0.32	26.60 ± 0.30	68.74 ± 0.36
VTLM + MMT *bilingual*	40.46 ± 0.64	81.58 ± 0.08	35.19 ± 0.16	79.79 ± 0.06	32.18 ± 0.21	75.94 ± 0.08
VGAMT *full bilingual*	43.30 ± 0.20	87.34 ± 0.08	38.30 ± 0.20	86.49 ± 0.07	35.70 ± 0.30	83.33 ± 0.08
VGAMT *SIGLIP-only multi.*	41.93 ± 0.75	85.79 ± 0.13	36.68 ± 0.23	84.72 ± 0.27	33.48 ± 0.13	81.05 ± 0.29
MMT – *cross-lingual zero-shot*						
M2KT-VPN *bilingual*	37.14 ± 0.69	75.51 ± 0.63	27.09 ± 0.40	72.29 ± 0.67	23.57 ± 0.54	66.07 ± 1.11
MMT – *zero-shot*						
Multilingual OpenFlamingo	28.86	82.31	23.91	80.91	21.99	76.58
ZeroMMT-600M (*ours*) *multi.*	36.22 ± 0.40	83.04 ± 0.39	33.11 ± 0.68	82.54 ± 0.17	29.04 ± 0.13	77.72 ± 0.16
ZeroMMT-1.3B (*ours*) *multi.*	37.63 ± 0.13	84.80 ± 0.19	36.24 ± 0.54	84.56 ± 0.19	31.66 ± 0.47	80.68 ± 0.14
ZeroMMT-3.3B (*ours*) *multi.*	39.58 ± 0.30	85.85 ± 0.05	37.97 ± 0.21	85.46 ± 0.13	33.71 ± 0.40	81.92 ± 0.16

Table 8: En→De results for Test2016, Test2017 and COCO subsets of M30K, avg. over 3 runs (± standard error).

	Test2016		Test2018	
	BLEU	COMET	BLEU	COMET
Text-only MT baselines				
NLLB-600M *distilled*	26.39	85.44	26.76	84.60
NLLB-1.3B	30.31	87.77	31.23	87.19
NLLB-3.3B	33.64	89.08	33.10	88.33
MMT – *fully supervised*				
Gated Fusion *bilingual*	30.80 ± 0.40	81.64 ± 0.32	25.80 ± 0.10	76.85 ± 0.18
VTLM + MMT *bilingual*	34.87 ± 0.19	84.15 ± 0.17	30.38 ± 0.35	80.64 ± 0.20
VGAMT *full bilingual*	37.60 ± 0.20	90.57 ± 0.08	34.20 ± 0.10	88.43 ± 0.06
VGAMT *SIGLIP-only multi.*	36.62 ± 0.42	88.63 ± 0.16	33.13 ± 0.23	86.28 ± 0.11
MMT – *cross-lingual zero-shot*				
M2KT-VPN *bilingual*	30.29 ± 0.54	81.33 ± 0.50	25.75 ± 0.28	75.93 ± 0.72
MMT – *zero-shot*				
Multilingual OpenFlamingo	3.22	71.27	3.31	70.18
ZeroMMT-600M (*ours*) *multi.*	25.66 ± 0.43	84.27 ± 0.36	24.82 ± 0.49	83.32 ± 0.14
ZeroMMT-1.3B (*ours*) *multi.*	29.98 ± 0.59	87.13 ± 0.27	30.29 ± 0.25	86.75 ± 0.28
ZeroMMT-3.3B (*ours*) *multi.*	32.99 ± 0.38	88.67 ± 0.07	33.08 ± 0.30	88.06 ± 0.11

Table 9: En→Cs results for Test2016 and Test2018 subsets of M30K, avg. over 3 runs (± standard error).

	EMMT		VATEX	
	BLEU	COMET	BLEU	COMET
Text-only MT baselines				
NLLB-600M *distilled*	14.72	53.60	17.42	62.03
NLLB-1.3B	18.42	56.51	18.02	63.54
NLLB-3.3B	21.01	57.77	20.09	64.77
MMT – *zero-shot*				
Multilingual OpenFlamingo	2.74	43.14	14.46	63.62
ZeroMMT-600M (*ours*) *multi.*	14.12 ± 0.09	52.39 ± 0.07	17.36 ± 0.13	61.82 ± 0.12
ZeroMMT-1.3B (*ours*) *multi.*	16.42 ± 0.15	54.84 ± 0.44	17.80 ± 0.16	63.50 ± 0.16
ZeroMMT-3.3B (*ours*) *multi.*	18.97 ± 0.59	56.34 ± 0.50	19.88 ± 0.25	64.87 ± 0.07

Table 10: En→Zh results for EMMT and VATEX test sets, averaged over 3 runs (± standard error).

γ	BLEU	COMET	CoMMuTE	BLEU	COMET	CoMMuTE
	ZeroMMT-600M			ZeroMMT-1.3B		
1.0	**32.73** ±12.33	**77.95** ±10.82	56.9 ±1.4	**35.62** ±12.58	**80.07** ±10.78	59.5 ±1.8
1.25	<u>32.39</u> ±12.24	<u>77.73</u> ±10.76	58.4 ±1.4	<u>35.25</u> ±12.56	<u>79.89</u> ±10.73	61.6 ±1.3
1.5	31.81 ±12.04	77.39 ±10.66	59.7 ±1.8	34.67 ±12.42	79.62 ±10.65	63.8 ±2.8
2.0	30.29 ±11.52	76.35 ±10.46	62.3 ±1.9	32.96 ±11.92	78.72 ±10.40	66.1 ±2.4
2.5	27.98 ±10.75	74.68 ±10.09	<u>64.1</u> ±2.1	30.36 ±11.29	77.09 ±10.04	<u>68.0</u> ±2.4
3.0	25.03 ±9.56	72.29 ±9.58	**65.4** ±2.2	27.06 ±10.21	74.60 ±9.49	**69.2** ±2.0

Table 11: Evolution of BLEU, COMET and CoMMuTE scores of our CFG-enabled ZeroMMT 600M and 1.3B models (aggregated over benchmarks and languages) compared to the vanilla, CFG-free ZeroMMT model (i.e. $\gamma = 1.0$). The best result of each model size is in **bold**. The second best result is <u>underlined</u>. See Section 7 for ZeroMMT-3.3B results.

	Test2016		Test2017		COCO		CoMMuTE
	BLEU	COMET	BLEU	COMET	BLEU	COMET	Accuracy
ZeroMMT-600M	48.62 ± 0.38	84.92 ± 0.09	**48.10** ± 0.11	**85.66** ± 0.16	**50.29** ± 0.82	83.78 ± 0.20	58.7 ±0.4
w/o VMLM	**49.01** ± 0.16	**85.09** ± 0.04	47.64 ± 0.19	85.59 ± 0.04	49.92 ± 0.28	**83.91** ± 0.02	50.0 ± 0.3
w/o KL	28.73 ± 5.97	78.51 ± 2.96	23.63 ± 5.99	76.95 ± 3.40	30.78 ± 6.48	76.50 ± 2.98	**60.4** ± 1.4
+ MMT w/o KL	48.68 ± 0.22	84.64 ± 0.21	47.64 ± 0.35	85.37 ± 0.05	49.40 ± 0.19	83.11 ± 0.18	56.8 ± 1.7

Table 12: Ablation study En→Fr. The best result is in **bold** and the second best result is underlined.

	Test2016		Test2017		COCO		CoMMuTE
	BLEU	COMET	BLEU	COMET	BLEU	COMET	Accuracy
ZeroMMT-600M	36.22 ± 0.40	83.04 ± 0.39	33.11 ± 0.68	82.54 ± 0.17	**29.04** ± 0.13	77.72 ± 0.16	55.7 ± 0.3
w/o VMLM	**37.17** ± 0.16	**83.59** ± 0.08	**33.72** ± 0.21	**83.10** ± 0.09	28.14 ± 0.38	**78.53** ± 0.21	50.0 ± 0.0
w/o KL	12.42 ± 5.76	67.10 ± 5.04	7.92 ± 4.36	64.92 ± 4.64	8.39 ± 4.17	61.32 ± 4.28	**56.8** ± 1.1
+ MMT w/o KL	35.95 ± 0.51	82.58 ± 0.10	32.72 ± 0.31	82.02 ± 0.11	27.40 ± 0.33	77.13 ± 0.02	54.6 ± 0.6

Table 13: Ablation study En→De. The best result is in **bold** and the second best result is underlined.

	Test2016		Test2018		CoMMuTE
	BLEU	COMET	BLEU	COMET	Accuracy
ZeroMMT-600M	25.66 ± 0.43	84.27 ± 0.36	24.82 ± 0.49	83.32 ± 0.14	55.5 ± 0.5
w/o VMLM	**26.49** ± 0.20	**85.17** ± 0.07	**26.55** ± 0.03	**84.34** ± 0.07	50.1 ± 0.2
w/o KL	10.90 ± 5.49	71.97 ± 5.42	8.15 ± 4.14	67.27 ± 5.68	**59.1** ± 0.8
+ MMT w/o KL	25.10 ± 0.27	83.78 ± 0.20	25.43 ± 0.10	82.66 ± 0.10	54.8 ± 0.9

Table 14: Ablation study En→Cs. The best result is in **bold** and the second best result is underlined.

	EMMT		VATEX		CoMMuTE
	BLEU	COMET	BLEU	COMET	Accuracy
ZeroMMT-600M	14.12 ± 0.09	52.39 ± 0.07	17.36 ± 0.13	61.82 ± 0.12	58.2 ± 1.1
w/o VMLM	**15.19** ± 0.27	**53.60** ± 0.11	**17.40** ± 0.10	61.95 ± 0.12	50.1 ± 0.2
w/o KL	1.12 ± 4.86	43.30 ± 1.06	8.99 ± 4.27	51.01 ± 5.38	**60.7** ± 0.5
+ MMT w/o KL	11.89 ± 0.44	51.56 ± 0.63	16.70 ± 0.18	**62.16** ± 0.24	56.5 ± 0.5

Table 15: Ablation study En→Zh. The best result is in **bold** and the second best result is underlined.

	Fr		De		Cs		CoMMuTE
	BLEU	COMET	BLEU	COMET	BLEU	COMET	
ZeroMMT-600M (*SIGLIP*)	**49.00** ±1.07	84.82 ±0.79	32.79 ±2.97	81.13 ±2.48	25.24 ±0.62	83.79 ±0.55	56.90 ±1.40
ZeroMMT-600M (*SIGLIP large*)	48.77 ±1.16	84.61 ±0.77	32.39 ±3.20	81.13 ±2.40	25.35 ±0.26	**83.89** ±0.40	56.73 ±1.65
ZeroMMT-600M (*CLIP*)	48.94 ±1.08	84.77 ±0.74	32.55 ±3.31	81.11 ±2.43	25.44 ±0.40	83.62 ±0.65	**57.25** ±1.70
ZeroMMT-600M (*VIT*)	48.86 ±1.45	84.64 ±0.67	32.62 ±3.03	81.17 ±2.23	25.10 ±0.37	83.64 ±0.53	55.59 ±0.96
ZeroMMT-600M (*ResNet*)	48.98 ±1.12	**84.90** ±0.71	**32.90** ±3.20	**81.34** ±2.41	**25.56** ±0.40	83.75 ±0.42	55.54 ±0.91

Table 16: Impact of visual features. Aggregated generation results for En→X. Fr and De results are averaged over Test2016, Test2017 from M30K and AmbiguousCOCO. Cs results are averaged over M30K Test2016 and Test2018. CoMMuTE results are averaged over languages. **Bold** is best result. Underline is second best.

	En	Source	:	A woman is folding sheets.
1	Fr	**Ours**	:	Une femme est en train de plier des feuilles.
		NLLB	:	Une femme est en train de plier des draps.
2	Fr	**Ours**	:	Une femme est en train de plier des draps.
		NLLB	:	Une femme est en train de plier des draps.

(a) French example from CoMMuTE. The English word 'sheets' can refer to 'paper' or 'bed sheets'. Correctly translated by ZeroMMT in both cases.

	En	Source	:	Give me the bill.
1	Fr	**Ours**	:	Donne-moi la facture.
		NLLB	:	Donnez-moi la facture.
2	Fr	**Ours**	:	Donne-moi le billet.
		NLLB	:	Donnez-moi la facture.

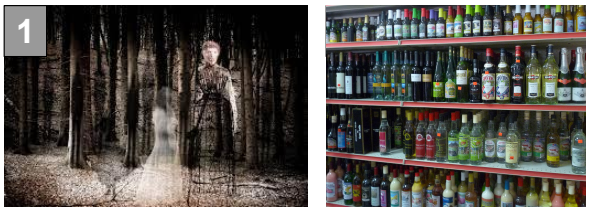
(b) French example from CoMMuTE. The English word 'bill' can refer to 'paper statement of money owed' or 'banknote'. Correctly translated by ZeroMMT in both cases.

	En	Source	:	Your gum doesn't look good!
1	De	**Ours**	:	Dein Zahnfleisch sieht nicht gut aus!
		NLLB	:	Dein Kaugummi sieht nicht gut aus!
2	De	**Ours**	:	Deine Kaugummi sieht nicht gut aus!
		NLLB	:	Dein Kaugummi sieht nicht gut aus!

(c) German example from CoMMuTE. The English word 'gum' can refer to 'mouth tissue' or 'chewing gum'. Correctly translated by ZeroMMT in both cases.

	En	Source	:	The match lasted a long time!
1	Ru	**Ours**	:	Матч длился очень долго!
		NLLB	:	Матч длился долго!
2	Ru	**Ours**	:	Спичка очень долго горел
		NLLB	:	Матч длился долго!

(d) Russian example from CoMMuTE. The English word 'match' can refer to 'sports game' or 'small flammable sticks'. Correctly translated by ZeroMMT in both cases.



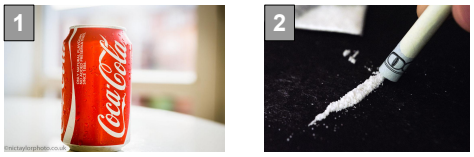
	En	Source	:	It must be the spirits!
1	Ar	**Ours**	:	ما من شك أنها الأرواح
		NLLB	:	لا بدّ أنّه الأرواح !
2	Ar	**Ours**	:	ما من شك أنها المشروبات الكحولية
		NLLB	:	لا بدّ أنّه الأرواح !

(e) Arabic example from CoMMuTE. The English word 'spirits' can refer to 'souls' or 'alcoholic beverages'. Correctly translated by ZeroMMT in both cases.

	En	Source	:	We've got another jam I'm afraid.
1	Zh	**Ours**	:	恐怕又卡纸了。
		NLLB	:	我们有另一个麻烦,我担心.
2	Zh	**Ours**	:	恐怕又遇上堵车了。
		NLLB	:	我们有另一个麻烦,我担心.

(f) Chinese example from CoMMuTE. The English word 'jam' can refer to 'paper stuck in a printer' or 'being stuck in the traffic'. Correctly translated by ZeroMMT in both cases.

Figure 9: Translations of CoMMuTE by our approach, ZeroMMT-600M, and the NLLB distilled MT model.



	En	Source	:	Have you got any coke?
1	Ru	**Ours**	:	У тебя есть кока-кола
		NLLB	:	У тебя есть кокаин?
2	Ru	**Ours**	:	У тебя есть кокаин?
		NLLB	:	У тебя есть кокаин?

(a) English–Russian example.

	En	Source	:	He is sitting on the trunk.
1	De	**Ours**	:	Er sitzt auf dem Baumstamm.
		NLLB	:	Er sitzt auf dem Kofferraum.
2	De	**Ours**	:	Er sitzt auf dem Kofferraum.
		NLLB	:	Er sitzt auf dem Kofferraum.

(b) English–German example.

	En	Source	:	Hand me the dough!
1	Ar	**Ours**	:	أعطني العجين!
		NLLB	:	أعطني النقود
2	Ar	**Ours**	:	أعطني النقود
		NLLB	:	أعطني النقود

(c) English–Arabic example.

	En	Source	:	We changed coaches.
1	Cs	**Ours**	:	Změnili jsme trenéry.
		NLLB	:	Změnili jsme trenéry.
2	Cs	**Ours**	:	Změnili jsme autobusy.
		NLLB	:	Změnili jsme trenéry.

(d) English–Czech example.

Figure 10: Additional translations of CoMMuTE by our approach, ZeroMMT-600M, and the NLLB distilled MT model.