CLIPTrans: Transferring Visual Knowledge with Pre-trained Models for Multimodal Machine Translation
Supplementary Material

Devaansh Gupta[1,2,*]
guptadm@bc.edu

Siddhant Kharbanda[3]
skharbanda@microsoft.com

Jiawei Zhou[4]
jzhou02@g.harvard.edu

Wanhua Li[4]
wanhua@seas.harvard.edu

Hanspeter Pfister[4]
pfister@seas.harvard.edu

Donglai Wei[1]
weidf@bc.edu

[1]Boston College [2]BITS Pilani [3]Microsoft India [4]Harvard University

S-1. Language Codes

The MT language codes mentioned in the paper along with their languages have been shown in Tab. S-1.

Code	Language	Code	Language
EN	English	ES	Spanish
DE	German	RO	Romanian
FR	French	AF	Afrikaans
CS	Czech		

Table S-1: Conventional MT Language codes.

S-2. Datasets

S-2.1. Details

Multi30k. Multi30k contains images sourced from the Flickr30k dataset [15] with English captions, professionally translated to German and extended to French and Czech. Conventionally, previous MMT methods have reported results only on the German and French splits. The test datasets involve Test2016 and Test2017 which were proposed in their respective years, along with the MSCOCO test set which contains 461 challenging out-of-domain instances from the MSCOCO dataset with ambiguous verbs.

WIT. WIT is sourced from Wikipedia images and their descriptions in multiple languages. We use this dataset to demonstrate results on low-resource and non-english language splits, specifically on EN $\rightarrow$ {RO, AF}, DE $\rightarrow$ ES and ES $\rightarrow$ FR. Apart from this, WIT also contains high-resource splits for EN $\rightarrow$ {DE, FR, ES}. These are annotated differently from Multi30k, since the descriptions are independently written for each image, thus inherently introducing noise in the paired translation data and increasing the dependence on images. We use the exact splits as proposed in [5] to ensure uniformity. Note that there can however be some variation in our scores since some images in the training data could not be downloaded. This does not affect the test set due to our text-only setting during inference.

Whenever needed, we apply preprocessing for both datasets following the input data format of respective pre-trained models.

S-2.2. Licences

All datasets used in this work are publicly available. WIT[1] [10] is available under the CC BY-SA 3.0 license. The license for Multi30k[2] [4] is unknown. Use of images from Flickr30k[3] are subject to Flickr Terms of Use[4].

S-3. Hyperparameters

Architectural Details. We combine two pre-trained models. M-CLIP [1] and mBART [11] to develop a multimodal multilingual model. mBART is initialized with its unsupervised pre-trained weights.[5] For M-CLIP we use the model variant consisting of an XLM-Roberta-Large[6] text encoder and a CLIP-ViT-B/32 [7] image encoder. The specific configurations of these models is shown in Tab. S-3.

Choice of Captioning Language. In the main paper, we demonstrate how captioning on multiple languages harms

[*]Work done as an intern at Boston College

[1]https://github.com/JerryYLi/valhalla-nmt/releases/tag/v0.1-datasets
[2]https://github.com/multi30k/dataset
[3]http://hockenmaier.cs.illinois.edu/DenotationGraph/
[4]https://www.flickr.com/help/terms/
[5]https://huggingface.co/facebook/mbart-large-50
[6]https://github.com/FreddeFrallan/Multilingual-CLIP
[7]https://huggingface.co/openai/clip-vit-base-patch32

# samples	Multi30k		WIT			
	EN → DE	EN → FR	EN → RO	EN → AF	DE → ES	ES → FR
Train	29k	29k	40k	18k	133k	122k
Validation	1k	1k	5k	5k	10k	10k
Test	2.5k	2.5k	1k	1k	2k	2k

Table S-2: Dataset statistics for Multi30k and WIT

	# Layers	# Attention Heads	Vocab/Patch Size	Embedding Dim	Feedforward Dim	Projection Dim
mBART	12	16	250k	1024	2048	-
XLM-Roberta-Large	24	12	250k	1024	4096	512
ViT-B/32	12	12	32	768	3072	512

Table S-3: Model statistics for CLIPTrans

Model	Multi30k				WIT		
	EN → DE						
	Test2016	Test2017	MSCOCO	Average	EN → RO	EN → AF	Average
CLIPTrans (Ours)	43.87	37.22	34.49		18.34	17.34	
Mapping Network Architectures							
CLIPTrans-MLP	41.94	35.96	33.35	-1.43	Unstable	10.49	-6.85
CLIPTrans-Enc	42.29	36.75	35.41	-0.37	17.86	17.54	-0.13
Injection of M-CLIP Embeddings							
Before <eos>	43.15	38.14	34.59	-0.10	17.45	16.97	-0.63

Table S-4: Additional Ablations on the Multi30k and WIT dataset

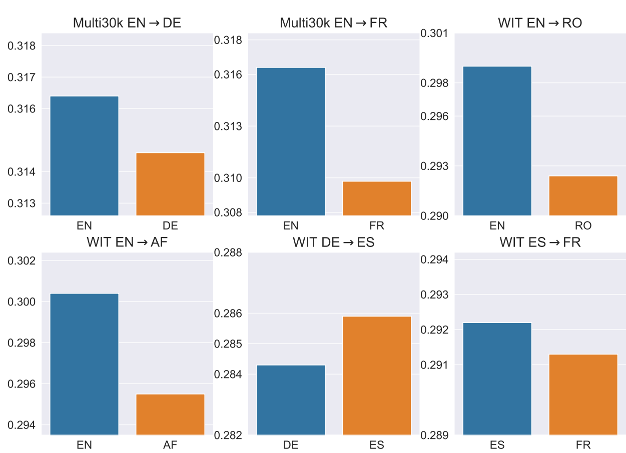

Figure S-1: Image-caption alignment of all the considered language pairs in their respective training splits. For each split, we perform captioning only on the language with higher similarity.

the performance of the mapping network. Therefore, during the first stage, we perform image captioning using a single language which is chosen on the basis of the image-caption alignment of that language on the training set with M-CLIP. This is calculated by finding the mean cosine similarity of the images and their captions in the M-CLIP encoding space across the training set. A summary of this is shown in Fig. S-1.

S-4. Additional Experiments

Dependence on Mapping Network Architecture. We have chosen the simplest mapping network for our main results, however, we also demonstrate variations of the same by training two additional models with identical hyperparameters – CLIPTrans-MLP and CLIPTrans-Enc. CLIPTrans-MLP employs fan MLP mapping network with the configuration as Linear→ReLU→Linear→PReLU. CLIPTrans-Enc projects the M-CLIP embedding to the required size, then applies a single transformer layer with two self-attention heads. The results of both are shown in Tab. S-4. While it may be possible to improve (or stabilize) these results via subsequent hyperparameter tuning, choosing a simple mapping network for CLIPTrans, enables us to set a lower bound on the results.

Injection of M-CLIP embeddings into mBART. During pre-training, the first token in the mBART decoder is the <eos> token which has the <bos> token as its label. To prevent misalignment with this design choice, we place the prefix sequence after this token. We ablate this and experiment by placing the prefix tokens before it or at the end

MMT Model	Inference	EN → DE			EN → FR			Average
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
Gumbel-Attention [8]	L+I	57.8	51.2	46.0	-	-	-	-13.97
CAP-ALL [6]		57.5	52.2	46.4	74.3	68.6	62.6	-11.40
GMNMT [14]		57.6	51.9	47.6	74.9	68.6	62.6	-11.13
DCCN [7]		56.8	49.9	45.7	76.4	70.3	65.0	-10.98
Gated Fusion* [13]		67.8	61.9	56.1	81.0	76.3	70.5	-2.73
ImagiT [9]	L	55.7	52.4	48.8	74.0	68.3	65.0	-10.97
RMMT* [13]		68.0	61.7	56.3	81.3	76.1	70.2	-2.73
VALHALLA [5]		68.8	62.5	57.0	81.4	76.4	70.9	-2.17
VALHALLA* [5]		69.3	62.8	57.5	81.8	77.1	71.4	-1.68
CLIPTrans (Ours)		**70.22**	**65.43**	**61.26**	**82.48**	**77.82**	**72.78**	

Table S-5: METEOR scores on the Multi30k dataset. Here we let * represent ensembled models. L+I represents both language and image are used during inference while L means only text is used during inference. **Bold** represents the highest score. We see CLIPTrans outperforms state-of-the-art methods across all settings.

Model	Under-Resourced		Non-English		Average
	EN → RO	EN → AF	DE → ES	ES → FR	
RMMT [13]	23.6	29.6	33.2	36.5	-4.79
UVR-NMT [16]	28.0	32.8	32.7	37.2	-2.84
VALHALLA [5]	30.4	34.2	34.3	37.5	-1.41
CLIPTrans (Ours)	**34.36**	**35.74**	**34.21**	**37.73**	

Table S-6: METEOR scores on the WIT dataset. We observe our method attains the best scores with a substantial margin.

of the sequence. Subsequently, the decoder self-attention mask is modified. As expected, we notice a slight drop in performance by placing them at the start. Placing at the end causes unstable training for all languages, which can be attributed to the lack of extra self-attention operations undergone by the prefix tokens as compared to placing them at the start, thus preventing them from properly adapting to the mBART.

METEOR. We show the METEOR [3] scores on the Multi30k dataset in Tab. S-5 and on WIT in Tab. S-6. Notably, CLIPTrans outperforms all previous SOTAs on METEOR as well.

Additional Results. In order to demonstrate the effectiveness of CLIPTrans for sentences outside the domain of the CLIP pre-training data, we evaluate on WMT2014 for EN→DE, FR. Following the undersampled settings in [5], we take a 100k random subset. Due to the lack of images, we only train stage 2 of CLIPTrans. As can be seen in Tab. **??**, we outperform the baseline across both languages.
For completeness, we also show results in Tab. **??** the EN → CS split of Multi30k, and note that we beat the mBART baseline.

S-5. Limitations

A potential limitation of our method is the computational cost associated with training larger pre-trained models. However, our method is general enough to be replicated on smaller or distilled models as well. Further, in order to take advantages of pre-trained weights, it is limited to the languages used in the pre-training data for M-CLIP and mBART. While this can be counteracted via zero-shot cross-lingual transfer approaches [2, 12], we leave that for discussion in future works.

Model	Multi30k(EN → CS)		WMT	
	Test2016	Test2018	EN → DE	EN → FR
mBART	35.20	32.02	19.58	29.35
CLIPTrans	**36.05**	**32.53**	**21.02**	**30.34**

Table S-7: Additional results on WMT and the EN → CS split of Multi30k.

S-6. Broader Impact

CLIPTrans can effectively ground images in multiple languages without requiring expensive post-pretraining steps and demonstrates how to effectively leverage exisiting pre-trained models in MMT. Beyond MMT, it can be considered as a generalized approach for developing better multimodal multilingual models using monolingual image captioning data which is of great practical importance. While negative impacts of this are hard to predict, it suffers from the same dataset and societal biases faced by vision and language models. While extensive work is being done to mitigate this, it is beyond the scope of this paper.

References

[1] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, June 2022. European Language Resources Association. S-1

[2] Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. *arXiv preprint arXiv:2104.08757*, 2021. S-3

[3] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. S-3

[4] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. S-1

[5] Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5216–5226, 2022. S-1, S-3

[6] Zhifeng Li, Yu Hong, Yuchen Pan, Jian Tang, Jianmin Yao, and Guodong Zhou. Feature-level incongruence reduction for multimodal translation. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 1–10, 2021. S-3

[7] Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329, 2020. S-3

[8] Pengbo Liu, Hailong Cao, and Tiejun Zhao. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*, 2021. S-3

[9] Quanyu Long, Mingxuan Wang, and Lei Li. Generative imagination elevates machine translation. *arXiv preprint arXiv:2009.09654*, 2020. S-3

[10] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. S-1

[11] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, 2021. S-1

[12] Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. Cross-lingual retrieval for iterative self-supervised training. *Advances in Neural Information Processing Systems*, 33:2207–2219, 2020. S-3

[13] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *arXiv preprint arXiv:2105.14462*, 2021. S-3

[14] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. *arXiv preprint arXiv:2007.08742*, 2020. S-3

[15] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. S-1

[16] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2020. S-3