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

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

Abstract

There has been a growing interest in developing multimodal machine translation (MMT) systems that enhance neural machine translation (NMT) with visual knowledge. This problem setup involves using images as auxiliary information during training, and more recently, eliminating their use during inference. Towards this end, previous works face a challenge in training powerful MMT models from scratch due to the scarcity of annotated multilingual vision-language data, especially for low-resource languages. Simultaneously, there has been an influx of multilingual pre-trained models for NMT and multimodal pre-trained models for vision-language tasks, primarily in English, which have shown exceptional generalisation ability. However, these are not directly applicable to MMT since they do not provide aligned multimodal multilingual features for generative tasks. To alleviate this issue, instead of designing complex modules for MMT, we propose CLIPTrans, which simply adapts the independently pre-trained multimodal M-CLIP and the multilingual mBART. In order to align their embedding spaces, mBART is conditioned on the M-CLIP features by a prefix sequence generated through a lightweight mapping network. We train this in a two-stage pipeline which warms up the model with image captioning before the actual translation task. Through experiments, we demonstrate the merits of this framework and consequently push forward the state-of-the-art across standard benchmarks by an average of ***+2.67 BLEU****. The code can be found at* www.github.com/devaansh100/CLIPTrans.

1. Introduction

Over the decades, Machine Translation (MT) has evolved from being rule-based [45], to more intricate probabilistic models [42, 14, 38, 23] and recently to end-to-end deep neural networks [1, 11, 62, 59] giving rise to the subdomain of Neural Machine Translation (NMT). Most recent NMT models largely rely on paired textual data and typically make use of transformer-based encoder-decoder models [62, 28] to set impressive benchmarks [35, 49]. With advancements in the transformer's ability to encode both images and texts in the same latent space [56, 26, 17, 39], there has been a rise in works [33, 34, 66, 54] leveraging images as auxiliary information to provide visual grounding to the translation task to enhance MT systems, a setting known as Multimodal Machine Translation (MMT).

For incorporation of the visual input, previous works have employed specifically engineered encoder-decoder architectures with multimodal attention modules [33, 34, 7, 71, 32, 75] that need to be learned from scratch. Consequently, they are forced to balance vision-language alignment with the translation task. Furthermore, to reduce the

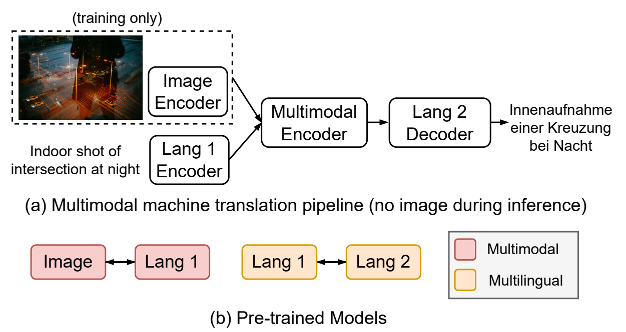


Figure 1: (a) Multimodal machine translation (MMT) models are hard to train due to the scarcity of triplet data, especially for low-resource languages. (b) Our work aims to leverage existing non-triplet pre-trained models for the MMT task (without image during inference setting).

*Work done as an intern at Boston College dependence of MMT on images during inference, previous works typically adopt one of two approaches where they either learn a *hallucination* network to generate image features from text [30, 37], or use retrieval modules to fetch one or more relevant images [73]. The former requires specially designed losses and difficult optimization while the latter comes with an extra computational cost at test time.

With an increase in popularity of transfer learning methods that make use of task-specific pre-trained unsupervised models, recent NMT works have observed a paradigm shift. However, a similar trend has not been witnessed in the MMT domain due to the requirement of data in the form of triplets comprising images and their bilingual captions, which limits transfer learning for three reasons: (i) pre-trained models for NMT are only trained on textual data [12, 13, 61, 68] (ii) existing pre-trained models are either multimodal with English as the only language [26, 56, 51, 60] or lack decoders for sequence generation [9, 22] (iii) MMT will require a multilingual multimodal network, which is difficult to train since triplets are expensive to source at the required scale, and existing triplet datasets cannot cover low-resource languages [31].

In this work, we aim to overcome these limitations and simplify the multimodal translation task by employing two independent pre-trained models as aforementioned in (i) and (ii). More specifically, we make use of M-CLIP [9] – a multilingual variant of the pre-trained multimodal CLIP [51] encoder – in an optimal training pipeline that tactfully enriches mBART [36] – a pre-trained text-only translation model – with powerful and well-aligned multimodal features. CLIP consists of visual and textual encoders that are trained on a large image-captioning dataset using contrastive learning which endows it with generalized, transferable representations for a variety of multimodal tasks [44, 40, 41, 27]. When provided with a text input at test time, M-CLIP essentially acts as a *hallucination* network by providing text embeddings pre-aligned with its visual counterpart. This not only removes the constraint of requiring images during inference but also inherently eliminates the need for hand-engineered architectures with complex training objectives aimed at vision-language alignment [20, 58]. Specifically, we employ a mapping network to transfer M-CLIP embeddings as decoder prefix to mBART and train the mBART decoder using a novel two-stage learning pipeline. In the first stage, we train the mBART decoder for the image-captioning task using a visual-textual decoder prefix sequence computed by a simple, lightweight mapping network from the M-CLIP image encoder. In stage two, the mBART decoder is trained for the translation task, generating decoder prefixes via the M-CLIP text encoder. Interestingly, this mimics the dataset annotation procedure for MMT datasets which first captions an image, then translates the caption while ensuring visual grounding with the image [54, 3]. Doing so enables transferring visual representations to the multilingual space, while effectively adapting the mBART attention maps to the newly introduced embeddings.

Contributions. (1) We present an architecture, CLIPTrans, that can capitalize on existing pre-trained LMs and multimodal models, thus simplifying the MMT pipeline by eliminating the use of specialized structures and intractable training objectives. (2) We propose a novel transfer-learning approach through a two-stage training pipeline wherein the first stage is a shared captioning task and the second is the translation task. We believe we are one of the first works to showcase the merits of using image captioning for adapting pre-trained models for MMT through a thorough analysis and demonstration of quantitative and qualitative results. (3) We surpass the previous state-of-the-art on MMT across two benchmarks by an average of **+2.88 BLEU**, and an average of **+3.64 BLEU** for under-resourced languages, without using images at test time, which further broadens the applicability of our method.

2. Related Work

Multimodal Machine Translation. MMT has been examined through various lenses [54, 33, 32, 73, 30, 4, 20, 7, 34, 71, 75, 6, 69, 53, 57, 21, 48], with the focus shifting from earlier works on RNN-based encoder-decoder networks to the recently proposed transformer architecture. As discussed earlier, fusion was done through special attention modules. Calixto *et al*. [7] introduces the use of spatial-visual image features through a doubly-attentive attention module and Calixto *et al*. and Liu *et al*. [6] further builds upon that to using global visual feature tokens in the source sentence. LIUM-CVC [4], MeMad [19] and DCCN [33] use an image-context reweighing of predicted token probabilities while decoding. Gated Fusion [66] and UVR-NMT [73] use an image-guided gating mechanism to incorporate image features in decoder cross-attention. In addition to this, UVR-NMT, like RMMT [66] also employs a retrieval module to fetch images during inference. Finally, VALHALLA [30] trains a multimodal encoder and visual hallucination module from scratch for MMT. With respect to using pre-trained models, Kong & Fan [70] adds a decoding head on top of a BERT model an expensive perform vision-language pre-training, similar to that of VisualBERT [26]. As an alternative to using pre-trained weights, GMNMT [71] incorporates a visually grounded multimodal graph built with BERT features into its training data.

Vision-Language Training. Combining vision and language has a long-standing research history. Learning generic cross-modal representations benefits various downstream tasks such as visual grounding [74], visual question answering [18], visual reasoning [72], and visual understanding [29]. Inspired by the success of BERT [13],

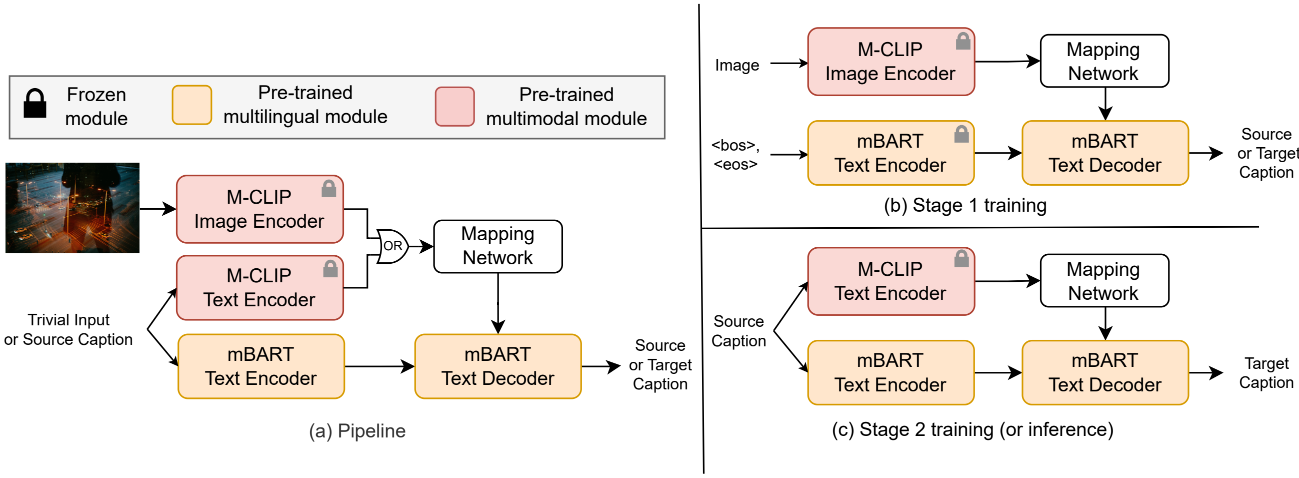


Figure 2: CLIPTrans framework overview. We show (a) all the modules in CLIPTrans and their wiring to enable transfer learning from pre-trained models for MMT. Along with that, we show the two-stage training pipeline with (b) the image captioning task in the first stage and (c) the language translation task in the second.

VisualBERT [26] and VL-BERT [56] take both visual and linguistic embedded features as input and train it on the Masked Language Modeling objective. VLMO [2] proposes a Mixture-of-Modality-Experts Transformer to unify vision-language training models which can process different modalities with a Transformer block. BEiT-3 [63] further extends it to a multi-way Transformer and attains state-of-the-art results on a broad range of benchmarks. While ClipCap [44] utilizes pre-trained GPT-2 and CLIP to obtain a lightweight image captioning model, BLIP [25] pre-trains language-image models by bootstrapping the captions. While these methods show strong generalization ability on various multimodal tasks, they need large vision-language paired datasets and focus on learning multimodal representations. In contrast, we are committed to image-free MMT during inference in a data-constrained setting.

There have been a plethora of works on transfer learning for machine translation [61, 52, 67, 43, 65, 64]. In this work, we propose a training pipeline, along with additional modules, for such models in order to leverage visual information during training to enhance text-only machine translation. More generally, our contribution to the research community can be summarised as a flexible method to enable multilingual generation from multimodal data for MMT, and subsequently to other multilingual seq2seq tasks which can benefit from images. While this is possible in works like PaLI [10], PaLM-E [15], it often cannot be finetuned on downstream data due to closed-source models and/or them being resource-intensive.

3. Method

Let $\mathcal{D}_v$ denote a vision-based multimodal corpus of image and text pairs (v, t), where v represents an image and t represents the corresponding text. Let $\mathcal{D}_l$ denote a language-based multilingual corpus of text and text pairs (x, y), where x represents a sentence in a source language and y represents its translation in a target language. In MMT, t is either aligned with x or y, thus creating a triplet data corpus consisting of (v, x, y) by combining $\mathcal{D}_v$ and $\mathcal{D}_l$. Our goal is to transfer the knowledge learned with the vision-language corpus $\mathcal{D}_v$ to augment the task of MT that is conducted on $\mathcal{D}_l$, with the effective fusion of pre-trained vision-language and language-only models.

3.1. Preliminaries

M-CLIP. Radford *et al.* [51] proposed the Contrastive Language-Image Pre-training (CLIP) encoders to align vision and language representations in a unified space. It is pre-trained on large-scale image-text paired corpus by matching text descriptions with images. In particular, the model consists of an image encoder $\text{Enc}_v^{\text{CLIP}}$ and a text encoder $\text{Enc}_t^{\text{CLIP}}$. Given an image-text pair (v, t), the encoded representations $\text{Enc}_v^{\text{CLIP}}(v)$ and $\text{Enc}_t^{\text{CLIP}}(t)$ are fix-sized vectors that are considered aligned with minimum cosine distance compared with the distances between unpaired texts with the same image. Although CLIP only works with English, a multilingual CLIP (M-CLIP) that extended the text encoder to work with different languages was also proposed [9]. We rely on the alignment structure of the vision-language representational space of M-CLIP to help transfer the knowledge learned with $\mathcal{D}_v$ to MT.

mBART. Pre-trained with sequence-to-sequence denoising objectives, BART (Bidirectional and Auto-Regressive Transformers) [24] is effective when fine-tuned with various text-to-text generation tasks including MT. It is composed of a Transformer text encoder $\text{Enc}_l^{\text{BART}}$ and a Transformer text decoder $\text{Dec}_l^{\text{BART}}$. Given a source sentence

$x = (x_1, x_2, \ldots, x_m)$, BART autoregressively generates the target sentence $y = (y_1, y_2, \ldots, y_n)$ through conditional language modeling

$$\begin{aligned} p(y|x) &= \prod_{i=1}^{n} p(y_i|y_{<i}, x) \\ &= \prod_{i=1}^{n} \mathrm{Dec}_l^{\mathrm{BART}}(y_{<i}, \mathrm{Enc}_l^{\mathrm{BART}}(x;\theta_e);\theta_d) \end{aligned} \tag{1}$$

where θ_e and θ_d are the parameters of the encoder and decoder, respectively, and source sentence x is first encoded by the encoder, and then utilized by the decoder along with the previously generated target $y_{<i}$ for predicting the next token y_i. Different attention mechanisms [62] are utilized in the decoder, with the source information $\mathrm{Enc}_l^{\mathrm{BART}}(x;\theta_e)$ passed through the cross-attention layers and the prefix information $y_{<i}$ passed through the self-attention layers with autoregressive masks. For the application of MT, multilingual BART (mBART) [36] that extends BART with pre-training on different languages achieves significant gains when fine-tuned for various MT tasks.

3.2. Vision and Language Integration

We aim to integrate vision and language information into a single framework by effectively fusing the multi-modal and multilingual pre-trained models, *i.e.* M-CLIP and mBART. We do so by applying a lightweight mapping network on the M-CLIP encoder representations to produce fixed-length embedding sequences as prefixes prepended to the mBART decoder input. In particular, we use a simple feedforward neural network for the mapping network, denoted as MN. Given an encoded M-CLIP representation vector $h \in \mathbb{R}^{d_c}$ either from image $h = \mathrm{Enc}_v^{\mathrm{CLIP}}(v)$ or from text $h = \mathrm{Enc}_t^{\mathrm{CLIP}}(t)$, it is mapped to a sequence of input embedding vectors for the mBART decoder:

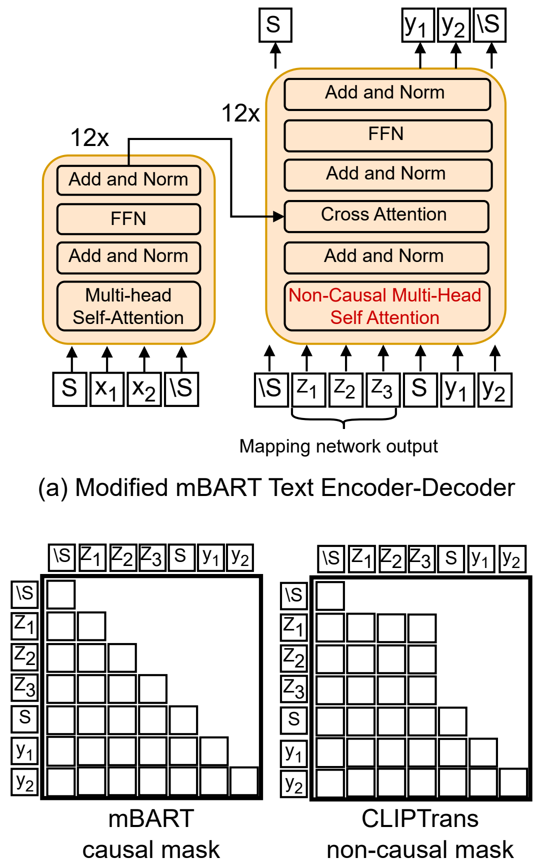
$$z = [z_1; z_2; \ldots; z_k] = \mathrm{MN}(h;\theta_m) \in \mathbb{R}^{k \cdot d_b} \tag{2}$$

where $[;]$ denotes vector concatenation, d_c is the visual-textual representation size from M-CLIP, d_b is the embedding size of the mBART decoder, k is the fixed length of the visual prefix embeddings, and θ_m is the learnable parameters of the mapping network.[1] Each $z_i \in \mathbb{R}^{d_b}$ for $i = 1, \ldots, k$ is serving as a visual-textual prefix token[2] to be utilized for the text generation with mBART:

$$\begin{aligned} p(y|h, x) &= \prod_{i=1}^{n} p(y_i|h, y_{<i}, x) \\ &= \prod_{i=1}^{n} \mathrm{Dec}_l^{\mathrm{BART}}([z, y_{<i}], \mathrm{Enc}_l^{\mathrm{BART}}(x;\theta_e);\theta_d) \end{aligned} \tag{3}$$

[1] We use a very light feedforward network with no hidden layers, with PReLU activation function on the output.

[2] This refers to mapped visual tokens, prepended to textual features.



(a) Modified mBART Text Encoder-Decoder

(b) Modifications in the mBART attention mask

Figure 3: (a) Detailed illustration of mBART in CLIPTrans, with modifications in the decoder while training. Note that x_i and y_i are tokens in the source and target language, respectively. $S, \backslash S$ are the special tokens $< bos >, < eos >$. Prefix tokens z_i are concatenated with the shifted output sequence before decoding. (b) The causal self-attention mask, which masks future tokens to ensure that the next token prediction is done only by attending to the previous ones, is modified to a non-causal one to enable bidirectional information flow amongst the visual context tokens.

Moreover, as the visual-textual prefix tokens z are produced all at once, we modify the mBART decoder self-attention mask to be bi-directional for the prefix segment. An illustration of our model is shown in Fig. 3.

3.3. Visual Knowledge Transfer Learning

Based on our integration of pre-trained M-CLIP and mBART, we propose a two-stage learning procedure that utilizes a vision-language corpus $\mathcal{D}_v$ and a language-only corpus $\mathcal{D}_l$ separately. The idea is to effectively utilize the internally aligned visual-textual representational structure of M-CLIP for transfer learning between images and texts.

Stage 1: Image-to-Text Captioning. The first stage is to warm up the mapping network MN and the mBART decoder $\mathrm{Dec}_l^{\mathrm{BART}}$ to utilize the visual information for text generation. Given an image-text pair (v, t), we transform the image into visual-textual prefix embeddings based on Eqn. 2 by first passing v into the M-CLIP image encoder, i.e. having $h = \mathrm{Enc}_v^{\mathrm{CLIP}}(v)$, and then applying the mapping network. We then learn to generate the text t from v based on the autoregressive process modeled by Eqn. 3 with mBART, where the target $y = t$, and the source is fixed at $x = (\texttt{<bos>}, \texttt{<eos>})$.[3] This is essentially an image captioning task, where the image information is encoded in the visual-textual prefix to the mBART decoder, and the caption is generated sequentially after the prefix. The mBART encoder does not provide any information with trivial x, which forces the model decoder to rely on the visual-textual prefix information for its generation. We only update the parameters of the mapping network and the mBART decoder (θ_m, θ_d) in this stage, and the M-CLIP and mBART encoders are kept frozen, as shown in Fig. 2.

Stage 2: Text-to-Text Translation. After stage 1 is done, we further tune the mapping network and the mBART model for the actual translation task relying on the paired textual corpus $\mathcal{D}_l$ without images. We swap out the M-CLIP image encoder with the M-CLIP text encoder directly for producing the visual-textual prefix embeddings. Specifically, with the translation paired sentences (x, y), we obtain the visual-textual prefix embeddings using Eqn. (2) again but with $h = \mathrm{Enc}_t^{\mathrm{CLIP}}(x)$. We then train the model with translation objectives to generate y from x based on Eqn. (3). Note that the source x is passed through both the mBART encoder and the M-CLIP text encoder to be utilized by the decoder for its generation. The parameters updated in this stage are the mapping network and mBART encoder and decoder, i.e. $(\theta_m, \theta_e, \theta_d)$. An illustration of this learning stage is shown in Fig. 2.

Note that the M-CLIP encoders are kept frozen in both stages. This ensures that its visual-textual representational space does not drift during training. We can utilize this structure to transfer the knowledge learned with visual input (stage 1) to the textual input (stage 2) in the form of the same decoder prefixes, as the visual and textual vectors encoded by M-CLIP are aligned during its pre-training. As a result, our training objectives are only the text generation cross-entropy loss in both stages,[4] without specially designed auxiliary losses to align the visual and textual information as required by previous approaches [20].

3.4. Inference

Our formulation in Eqn. 3 integrates M-CLIP encodings to help MT with the mBART encoder-decoder backbone. The visual-textual representations from M-CLIP allow different application scenarios for MT under our framework. When we have additional input of the image v and $h = \mathrm{Enc}_v^{\mathrm{CLIP}}(v)$, we can achieve vision-based MMT. For our basic application of text-only MT where we do not have additional image information during inference, we can simply set $h = \mathrm{Enc}_t^{\mathrm{CLIP}}(x)$ from the source sentence, similar to visual hallucination from the text during inference time [30]. Decoding can start after the visual-textual prefix computations, either through greedy search or beam search.

4. Experiments

4.1. Experimental Setup

Datasets. We demonstrate the effectiveness of our model on two public benchmarks: Multi30k [16] and Wikipedia Image Text (WIT) [55]. Multi30k is a widely used MMT benchmark which is a multilingual extension of the Flickr30k dataset that expands EN captions to DE and FR. Evaluation is performed on three standard test splits - Test2016, Test2017, and MSCOCO. MSCOCO test split consists of sentences with ambiguous verbs and out-of-domain data points from the COCO Captions dataset, which is considered a generally difficult setting for MMT models [66]. WIT is a multilingual dataset created by extracting image text pairs from Wikipedia in various languages. We use this dataset to set new benchmarks on non-English (DE $\rightarrow$ ES, ES $\rightarrow$ FR) and low-resource translations (EN $\rightarrow$ AF, RO). Additionally, results on WMT and the EN $\rightarrow$ CS are presented in the supplementary material.

Implementation details. Our models are trained using the previously discussed two-stage training pipeline. Each training stage is trained on 4 A100 GPUs using an AdamW optimizer and Polynomial Decaying Schedule for 15 epochs with a batch size of 256 and a learning rate set to 1e-5. Text decoding is done using beam search with a beam size of 5. All implementations are done in Pytorch using Huggingface Transformers. For the first stage, we pick either the source or target language for captioning depending on their training set alignment in M-CLIP.

Evaluation Metrics. All comparisons are made using BLEU [46], calculated with SacreBLEU [50], which is the gold standard for evaluating translation models. Unless otherwise mentioned, we report results using the checkpoint attaining the highest BLEU score on the validation set. We also benchmark our model on the METEOR metric, calculated with the evaluate library[5]. This can be found in the Supplementary Material.

[3] These are two special tokens marking the start and end of a sentence.

[4] No loss is computed on the visual-textual prefix embeddings.

[5] www.huggingface.co/spaces/evaluate-metric/meteor

MMT Model	Inference	EN → DE			EN → FR			Average
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
Gumbel-Attention [34]	L+I	39.20	31.40	26.90	-	-	-	-6.03
CAP-ALL [32]		39.60	33.00	27.60	60.10	52.80	44.30	-4.86
GMNMT [71]		39.80	32.20	28.70	60.90	53.90	-	-4.44
DCCN [33]		39.70	31.00	26.70	61.20	54.30	45.40	-4.71
Gated Fusion* [66]		42.00	33.60	29.00	61.70	54.80	44.90	-3.43
ImagiT [37]	L	38.50	32.10	28.70	59.70	52.40	45.30	-4.98
UVR-NMT [73]		36.90	28.60	-	58.30	48.70	-	-7.68
VMMT [8]		38.40	30.10	25.50	-	-	-	-7.19
IKD-MMT [47]		41.28	33.83	30.17	62.53	54.84	-	-5.02
RMMT* [66]		41.40	32.90	30.00	62.10	54.40	44.50	-3.54
VALHALLA [30]		41.90	34.00	30.30	62.30	55.10	45.70	-2.88
VALHALLA* [30]		42.70	35.10	30.70	63.10	56.00	46.50	-2.08
CLIPTrans (Ours)		**43.87**	**37.22**	**34.49**	**64.55**	**57.59**	**48.83**	

Table 1: Results 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 BLEU 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 [66]	9.90	9.80	11.00	15.90	-4.89
UVR-NMT [73]	12.50	11.60	10.90	16.40	-3.69
VALHALLA [30]	14.40	14.00	11.30	16.60	-2.46
CLIPTrans (Ours)	**18.34**	**17.34**	**13.06**	**17.41**	

Table 2: Results on the WIT dataset. We observe our method attains the best BLEU scores with a substantial margin.

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	
- Image Captioning	42.17	37.51	34.37	-0.51	17.99	16.30	-0.69
+ Multilingual Image Captioning	41.24	36.59	34.53	-1.07	17.76	15.87	-1.03
CLIPTrans-reg	43.40	36.44	34.67	-0.36	16.69	16.21	-1.39
+ Image Captioning	43.35	37.11	34.69	-0.14	17.76	17.65	-0.13
mBART	41.66	36.87	34.14	-0.97	14.87	15.21	-2.80
CLIPTrans (M)	43.40	36.44	34.67	-0.36	17.27	17.31	-0.55
CLIPTrans-SS	42.13	36.17	33.90	-1.12	17.84	16.36	-0.74
CLIPTrans-FT	42.79	36.92	34.10	-0.59	17.56	17.43	-0.34
CLIPTrans-CLIP	42.79	37.39	34.04	-1.90	18.31	17.21	-0.08

Table 3: Ablation Results on the Multi30k dataset and WIT dataset.

4.2. Benchmark Results

Results on Multi30K. As shown in Table 1, our method consistently outperforms all previous state-of-the-art methods and achieves the best BLEU scores across all language-test set splits. We compare our architecture with two kinds of methods: (i) conventional MMT methods that require images during inference and, (ii) methods that do not make use of images during inference. Numbers for comparison are directly quoted from the publication where possible or are obtained using their publicly available codebase.

Specifically, in comparison with the conventional MMT methods that require images during inference, we observe that our method attains +3.43 BLEU improvements on average over the Gated Fusion method [66]. These empirical gains validate our model's ability to effectively transfer visual knowledge from M-CLIP models for text-only test time translation. Next, in comparison with MMT approaches utilizing text-only input during inference, CLIPTrans significantly outperforms UVR-NMT [73] across all metrics without performing multiple image retrieval during inference. Notably, CLIPTrans outperforms not only the previous state-of-the-art method VALHALLA [30] by an average of +2.88 BLEU score without training a heavily-engineered hallucination transformer but also its ensemble by a significant margin using only a single instance. We attribute these improvements to using pre-trained weights, thus illustrating their effectiveness in MMT. We observe the highest gains on the difficult MSCOCO test split, which further validates the superiority of our training pipeline at effectively endowing the mBART decoder with visual information.

Results on WIT. Tab. 2 shows the comparison results on the WIT dataset. We observe that CLIPTrans consistently outperforms existing methods in both under-resourced and non-English settings. Compared with VALHALLA, our method attains +2.46 BLEU improvements on average, which illustrates the superiority of CLIPTrans over existing methods. Our method shows more significant performance gains on under-resourced settings, where CLIPTrans obtains 3.94 and 3.34 BLEU improvements on the EN $\rightarrow$ RO and EN $\rightarrow$ AF tasks, respectively. Relatively smaller gains were seen on non-English benchmarks which can be attributed to two factors (i) there is an English-centric bias in WIT due to which the images are not very well aligned for non-English pairs, as argued in [30] and, (ii) imperfect alignment of M-CLIP image-text embeddings for non-English languages since, during training, their representations are derived by machine translating English text to the target language which may introduce inaccuracies.

4.3. Ablation & Analysis

We ablate our training pipeline on both datasets on three language pairs, EN $\rightarrow$ {DE, RO, AF}, as shown in Tab. 3.

mBART. We introduce a new baseline to directly compare the effect of introducing M-CLIP embeddings. Thus, we train a text-only mBART on multilingual captions of each language pair independently. As can be seen in Tab. 3, mBART is an extremely strong baseline, since it even outperforms ensembles of previous MMT SOTAs on Multi30k and parallels them on WIT. Adding M-CLIP embeddings in CLIPTrans consistently improves upon this baseline, showing the advantage of fusing pre-trained models.

Effect of Image Captioning. In order to understand the benefit of the image-to-text captioning stage on CLIPTrans, we directly train on translation without the first stage training and report its scores. The performance drops by an average of 0.6 BLEU which shows that captioning is essential for translation in our pipeline and serves as an effective warm-up strategy for the decoder and the mapping network.

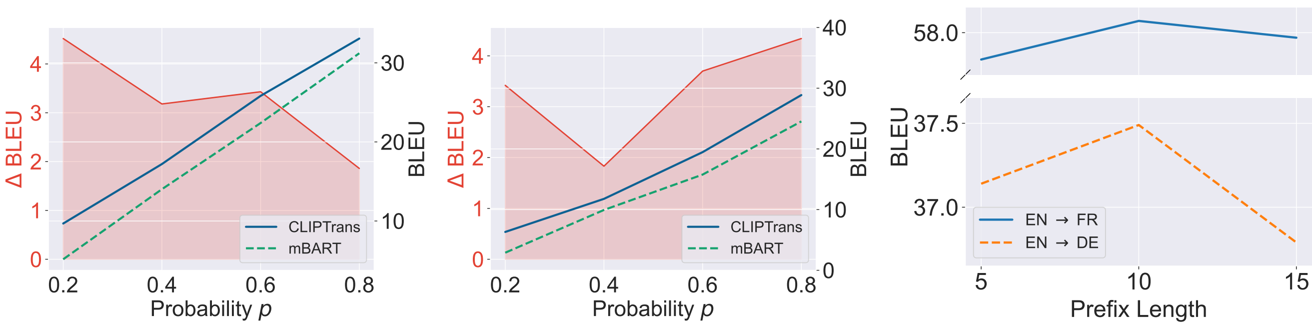
Choice of Captioning Language. As mentioned before, we caption on only one language between the source and target, depending on their alignment in M-CLIP. To validate this, we also train our model on both languages (+ multilingual image captioning in Tab. 3) and notice a performance drop - we conjecture this is because both languages influence the gradients in different directions, thus leading to sub-optimal learning.

CLIPTrans in Traditional MMT Pipelines. Previous MMT approaches [7, 66, 33] used a simple training pipeline where the image and its caption were supplied as inputs and the translated caption as a target. To demonstrate the generalizability of our architecture, we train CLIPTrans-reg under this setting by passing the image through the M-CLIP Image encoder and the source caption through the mBART encoder. While we notice a drop in performance, more significantly on WIT, we still outperform previous SOTAs. Furthermore, warming up the weights with image captioning brings CLIPTrans-reg closer to CLIPTrans, thus validating both, the superiority of our transfer-learning approach and the importance of image captioning for alignment.

Using Ground-Truth Images in inference. In order to ensure that we perform accurate hallucination during inference, we replace the M-CLIP text encoder with the M-CLIP image encoder and use ground truth images(CLIPTrans(M)). We notice a slight drop in performance, as shown in Fig. 3 since this introduces a train-test disparity, as discussed in [30]. We further note this disparity when CLIPTrans-reg is trained with image captioning and tested only with text. Thus our approach effectively mitigates this issue without the use of auxiliary losses.

Need for Visual Context. As noted by [5, 66], images often act as regularizers, especially on the Multi30k dataset, due to the high quality of the paired translation data. They further study the effect of degrading inputs during training and inference, since this would force the model to attend to the images. We believe our image captioning stage enables that and thus demonstrates the ability of CLIPTrans to recover translations, when compared to an mBART trained under the same scenario in Fig. 4. For this experiment, we randomly drop tokens from the train and test set with a probability p. Furthermore, CLIPTrans uses ground truth images during stage 2 training and inference to study their necessity. While the trends with respect to p are dataset dependent, we consistently see an improvement in CLIPTrans by an average of +3.3 BLEU, even for the low-masking scenario, thus showing the ability of mBART to effectively adapt and utilize the visual context.

Sensitivity to Prefix Length:. We ablate the sensitivity to the prefix length and note that our performance peaks at a length of 10, as shown in Fig. 4c. We believe that reducing the prefix length prevents the prefix sequence from being expressive enough while increasing it adds redundancies.



(a) (b) (c)

Figure 4: Evaluation on noisy inputs on CLIPTrans and mBART on the (a) Test2016 and (b) Test2017 split of the Multi30k dataset on EN $\rightarrow$ DE. Recovery is consistently higher than that of an mBART. (c) Sensitivity to the prefix length of different language pairs on the Test2017 split of the Multi30k dataset.

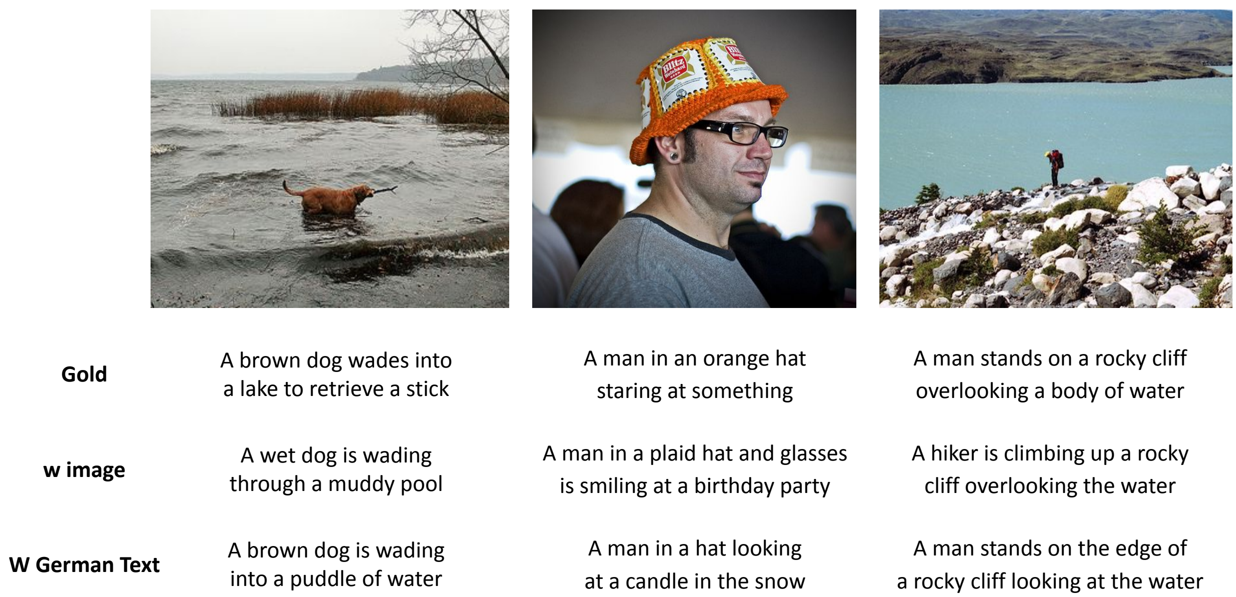


Figure 5: Qualitative results of CLIPTrans after the captioning stage, on both captioning and zero-shot German to English translation. Data points are from the Test2016 test set of Multi30k. As is visible, CLIP tokens are coherently decoded by the mBART into captions and zero-shot translations.

Training Pipelines. We ablate our training pipeline with two variations: (i) single-stage training where we perform stage 1 and stage 2 together. This is done by backpropagating on one data point twice in a batch - once with the image and once with the source text. As shown in Tab. 3, CLIPTrans-SS gives inferior results than our proposed two-stage pipeline. (ii) Since we no longer need the CLIP image and text encoder to be aligned after Stage 1, we try fine-tuning the CLIP text encoder in Stage 2. As can be seen in Tab. 3, CLIPTrans-FT also attains lower scores. We believe this happens since simultaneous optimization of the mBART encoder, decoder, and CLIP might be difficult.

Choice of image-text encoder. We experiment with different multimodal encoders in our pipeline. Given our setup, having a pre-aligned image-text encoder is imperative. Hence we choose CLIP as presented in [51] for this experiment and train CLIPTrans-CLIP in Tab. 3. Note that M-CLIP uses the same visual encoder as CLIP. Moreover, it facilitates Non-English translations which is not possible with other English-only models. A slight drop in performance shows that M-CLIP's multilingual pre-training creates better language features, even for English[6].

[6]The sharp average drop on Multi30k in Tab. 3 is largely due to the score on the Test2016 split. We do not see such variations with WIT.

Gold	A boston terrier is running on lush green grass in front of a white fence	Five people wearing winter jackets stand in the snow, with snowmobiles in the background	A guy works on a building
Input	A <mask> is running on lush green grass in front of a white fence	Five people wearing <mask> stand in the snow, with snowmobiles in the background	A guy works on a <mask>
Ours	Ein Hund läuft vor einem weißen Zaun auf leuchtendem grünem Gras *(A **dog** runs in front of a white fence on bright green grass)*	Fünf Personen in Schneekleidung stehen im Schnee *(Five people in **snow clothes** stand in the snow)*	Ein Mann arbeitet an einem Gebäude *(A guy works on a **building**)*
mBART	Ein Lächelndes grünes Gras vor einem weißen Zaun *(A **smiling** green grass in front of a white fence)*	Vier Personen, die Schnee tragen, stehen im Schnee *(Four people **carrying snow** stand in the snow)*	Ein Mann arbeitet an einer Maschine *(A guy works on a **Machine**)*

Figure 6: Qualitative results of CLIPTrans on recovery of visually grounded masked tokens, when compared to an mBART. Data points are from the Test2016 test set of Multi30k. The gold sentence is the ground truth. The *italicized* sentence in the bracket shows the English translation of the German Text obtained via Google Translate, and **bold** shows the predicted masked word. The image context is effectively utilized and the predicted words are not solely a consequence of the language model, as demonstrated by the mBART translations.

Qualitative Results. In order to further ensure that our results are not solely achieved due to regularisation and that M-CLIP embeddings are not being treated as noise, we show qualitative results that the decoder can actually derive coherent information from them. This is done by using CLIPTrans to decode M-CLIP tokens when no extra information is provided by the mBART encoder. Image captioning results can be seen in Fig. 5.

We replace the M-CLIP image encoder with its text encoder and evaluate zero-shot translations, by using German text embeddings from M-CLIP to show that captioning knowledge can be transferred to translation due to the inherent structure of M-CLIP. Without extra information from the mBART encoder, we demonstrate in Fig. 5 that the decoder can understand a gist of what the translation should be, but does not use the fine-grained context. This context, along with the exact words to be used, is provided when we train with the mBART encoder in the second stage.

Finally, we also show qualitative results for recovery of masked token from the image in Fig. 6. This is done by masking visually grounded phrases in the source text and providing the ground truth image to recover the masked token. We also compare these results with an mBART to ensure that the recovery is a consequence of visual grounding and not a consequence of the language model. We note that while mBART ends up hallucinating phrases, CLIPTrans can recover the phrase accurately.

5. Conclusion

This work presents CLIPTrans, a versatile approach to enable leveraging independent pre-trained models, specifically the multimodal M-CLIP and multilingual mBART, for MMT without using heavily engineered architectures or any external data. Alongside, it presents a two-stage training pipeline wherein the first stage involves an image-to-text captioning task, and the second involves a text-to-text translation task. The efficacy of this schedule and the advantages of transfer learning through image captioning are thoroughly discussed and analyzed. Breaking down the problem as we do further allows us to naturally eliminate the constraint of images during inference without employing complex optimization strategies.

We not only push the state-of-the-art across multiple datasets, but also set strong text-only baselines with mBART that outperforms previous MMT SOTAs. Given the flexibility of our method, we believe our work could lead future works toward a relaxed MMT setting using unsupervised data. This will enable using existing large-scale datasets during training, thus pushing the domain forward and reducing the reliance on current small-scale datasets.

Acknowledgements

This research is supported in part by the NSF award IIS-2239688 and NIH award R01HD104969.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 3

[3] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, 2018. 2

[4] Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018. 2

[5] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *NAACL*, 2019. 7

[6] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*, 2017. 2

[7] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, 2017. 1, 2, 7

[8] Iacer Calixto, Miguel Rios, and Wilker Aziz. Latent variable model for multi-modal translation. In *ACL*, 2019. 6

[9] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022. 2, 3

[10] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014. 1

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[14] Yuan Ding and Martha Palmer. Machine translation using probabilistic synchronous dependency insertion grammars. In *ACL*, 2005. 1

[15] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[16] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 2016. 5

[17] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 1

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[19] Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018. 2

[20] Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *ACL*, 2020. 2, 5

[21] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016. 2

[22] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *NAACL*, 2021. 2

[23] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL companion volume proceedings of the demo and poster sessions*, 2007. 1

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. 3

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 2, 3

[27] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 2

[28] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2label: A language modeling framework for multi-attribute learning. In *ECCV*, 2022. 1
[29] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In *NeurIPS*, 2022. 2
[30] 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 *CVPR*, 2022. 2, 5, 6, 7
[31] Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *ACL*, 2023. 2
[32] 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*, 2021. 1, 2, 6
[33] 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 *ACM MM*, 2020. 1, 2, 6, 7
[34] Pengbo Liu, Hailong Cao, and Tiejun Zhao. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*, 2021. 1, 2, 6
[35] Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*, 2020. 1
[36] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *TACL*, 2020. 2, 4
[37] Quanyu Long, Mingxuan Wang, and Lei Li. Generative imagination elevates machine translation. In *NAACL: Human Language Technologies*, 2021. 2, 6
[38] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 2008. 1
[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019. 1
[40] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022. 2
[41] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 2022. 2
[42] Daniel Marcu and Daniel Wong. A phrase-based, joint probability model for statistical machine translation. In *EMNLP*, 2002. 1
[43] Sachin Mehta, Rik Koncel-Kedziorski, Mohammad Rastegari, and Hannaneh Hajishirzi. Define: Deep factorized input token embeddings for neural sequence modeling. In *ICLR*, 2019. 3
[44] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 3
[45] Eric Nyberg and Teruko Mitamura. The kant system: Fast, accurate, high-quality translation in practical domains. In *COLING*, 1992. 1
[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5
[47] Ru Peng, Yawen Zeng, and Jake Zhao. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation. In *EMNLP*, 2022. 6
[48] Ru Peng, Yawen Zeng, and Junbo Zhao. Hybridvocab: Towards multi-modal machine translation via multi-aspect alignment. In *ICMR*, 2022. 2
[49] Martin Popel. Cuni transformer neural mt system for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018. 1
[50] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018. 5
[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 8
[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 3
[53] Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *CVPR*, 2020. 2
[54] Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 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*, 2016. 1, 2
[55] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *ACM SIGIR*, 2021. 5
[56] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 2, 3
[57] Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation. In *CVPR*, 2019. 2
[58] Dídac Surís, Dave Epstein, and Carl Vondrick. Globetrotter: Unsupervised multilingual translation from visual alignment. *arXiv e-prints*, 2020. 2
[59] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *NeurIPS*, 2014. 1
[60] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2019. 2

[61] 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 ACL: IJCNLP*, 2021. 2, 3

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1, 4

[63] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, 2023. 3

[64] Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *ACL*, 2022. 3

[65] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022. 3

[66] 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. In *ACL: IJCNLP (Volume 1: Long Papers)*, 2021. 1, 2, 5, 6, 7

[67] Haoran Xu, Benjamin Van Durme, and Kenton Murray. BERT, mBERT, or BiBERT? a study on contextualized embeddings for neural machine translation. In *EMNLP*, 2021. 3

[68] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*, 2021. 2

[69] Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, and Vicente Ordonez. Using visual feature space as a pivot across languages. In *Findings of the ACL: EMNLP 2020*, 2020. 2

[70] Kong Yawei and Kai Fan. Probing multi-modal machine translation with pre-trained language model. In *Findings of ACL: IJCNLP*, 2021. 2

[71] 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. In *ACL*, 2020. 1, 2, 6

[72] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 2

[73] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *ICLR*, 2020. 2, 6, 7

[74] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NeurIPS*, 2020. 2

[75] Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multi-modal machine translation. In *EMNLP*, 2018. 1, 2