Visual Prediction Improves Zero-Shot Cross-Modal Machine Translation

Tosho Hirasawa[†]   **Emanuele Bugliarello**[⋆]
Desmond Elliott[⋆]   **Mamoru Komachi**[‡]
[†]Tokyo Metropolitan University
[⋆]Department of Computer Science, University of Copenhagen
[‡]Hitotsubashi University
hirasawa-tosho@ed.tmu.ac.jp

Abstract

Multimodal machine translation (MMT) systems have been successfully developed in recent years for a few language pairs. However, training such models usually requires tuples of a source language text, target language text, and images. Obtaining these data involves expensive human annotations, making it difficult to develop models for unseen text-only language pairs. In this work, we propose the task of **zero-shot cross-modal machine translation** aiming to transfer multimodal knowledge from an existing multimodal parallel corpus into a new translation direction. We also introduce a novel MMT model with a visual prediction network to learn visual features grounded on multimodal parallel data and provide pseudo-features for text-only language pairs. With this training paradigm, our MMT model outperforms its text-only counterpart. In our extensive analyses, we show that (i) the selection of visual features is important, and (ii) training on image-aware translations and being grounded on a similar language pair are mandatory. Our code are available at https://github.com/toshohirasawa/zeroshot-crossmodal-mt

1 Introduction

Multimodal machine translation (MMT) aims to improve translation quality with the help of other modalities, such as images (Specia et al., 2016) or videos (Wang et al., 2019). MMT models have shown promising improvement over their text-only neural machine translation (MT) counterparts, especially when it matters (Li et al., 2021; Lala and Specia, 2018; Gella et al., 2019). While prior work has successfully developed MMT models for language pairs with available multimodal parallel corpora, incorporating visual information into language pairs with no multimodal dataset has

Modality	Lang.	Examples
Text	> 700	BG, CS, DA, DE, EL, ES, ET, FR, JA, …
Text+Image	∼ 10	DE, FR, CS, JA, …

Table 1: Number of target languages with text-only (Text) or multimodal (Text+Image) parallel corpora for the translation from English.

received limited attention. As shown in Table 1, multimodal parallel corpora are only available for a few language pairs (Elliott et al., 2016, 2017; Barrault et al., 2018; Nakayama et al., 2020; Sanayai Meetei et al., 2019; Wang et al., 2019), which is quite less than the language pairs with text-only parallel corpora. Since building a multimodal parallel corpus by professional translators is costly and time-consuming (Wang et al., 2019), creating high-quality multimodal parallel corpora for many language pairs is not feasible.

One approach to addressing this problem is zero-shot cross-lingual transfer, which has proven successful in text-only machine translation (Firat et al., 2016; Johnson et al., 2017, *inter-alia*). In this paper, we investigate whether this success also extends to a multimodal setting. To this end, we propose the task of **zero-shot cross-modal machine translation**, where models need to perform multimodal machine translation in language pairs that lack multimodal parallel training data. In this task, there are still language pairs with multimodal training data, but the target language pairs consist of text-only training data.

To tackle this novel task, we propose a simple **M2KT-VPN** method that aims at performing Multimodal Knowledge Transfer via Visual Prediction Network in the *zero-shot cross-modal* translation setup. Inspired by El-

liott and Kádár (2017), a visual prediction network is employed to mimic visual features from the textual modality. We hypothesize that the predicted feature can help bridge the gap between text-only and multimodal translation pairs, so the model is not surprised when it receives true images at inference time.

The contributions of this work are as follows:

- We introduce a novel task, namely **zero-shot cross-modal machine translation** task, aiming to build MMT systems that can transfer multimodal knowledge from multimodal language pairs into text-only language pairs.

- We propose the **M2KT-VPN** model, a Transformer-based MMT model along with a visual prediction network, and show its zero-shot cross-modal translation capability.

- Our findings suggest the importance of image-aware translations and language similarity between translation directions.

2 Zero-shot Cross-Modal Machine Translation

We propose a new challenge for multimodal machine translation systems that we denote **zero-shot cross-modal machine translation** (Figure 1). This task is motivated by the real-world lack and cost of multimodal parallel corpora, which inhibits the development of multimodal translation systems beyond a few, mostly Indo-European, language pairs.

Task definition. The *zero-shot cross-modal* machine translation task aims to transfer multimodal knowledge learned from a (visually) grounded language pair into a language pair with no multimodal information at training. We define the two types of machine translation resources used for this task as follows:

- **Grounded language pairs**: language pairs where a multimodal parallel corpus is available, both at training and test time.

- **Zero-shot language pairs**: language pairs that only have a text parallel corpus for training, but have multimodal parallel data for test.
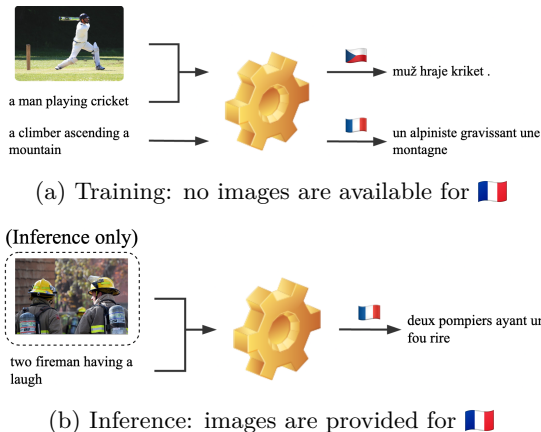
(a) Training: no images are available for 🇫🇷

(Inference only)

(b) Inference: images are provided for 🇫🇷

Figure 1: Overview of the **zero-shot cross-modal machine translation** task. For the zero-shot language pair (e.g., 🇫🇷), images are unavailable during training (a), but given at the inference (b).

Thus, a model is encouraged to transfer multimodal knowledge learned from grounded language pairs to zero-shot ones in order to best leverage multimodal data that may be available at test time.

Notation. We consider the following setup in our paper. Given a sequence of N tokens in a given source language, $\mathbf{x} = \langle x_1, x_2, \cdots, x_N \rangle$, and its associated image z, a multimodal machine translation model learns to translate $\mathbf{x}$ into a sentence of M tokens in a target language, $\mathbf{y} = \langle y_1, y_2, \cdots, y_M \rangle$. In the following, we directly consider a dense representation of the image z given by a visual feature extractor, which outputs I features that are then projected into a given model dimension d, $\mathbf{H}_z \in \mathbb{R}^{I \times d}$.

3 Proposed Approach: M2KT-VPN

In this section, we introduce a new MMT model, called **M2KT-VPN**, which aims to transfer multimodal knowledge learned from the multimodal corpus into the zero-shot language pair. M2KT-VPN comprises four modules (Figure 2): a Transformer (Vaswani et al., 2017) encoder to encode a source sentence, a visual prediction network (VPN) to predict a visual feature, a fusion module to incorporate multimodal information, and a Transformer decoder to generate a system output. All modules are trained simultaneously on grounded and zero-shot language pairs.
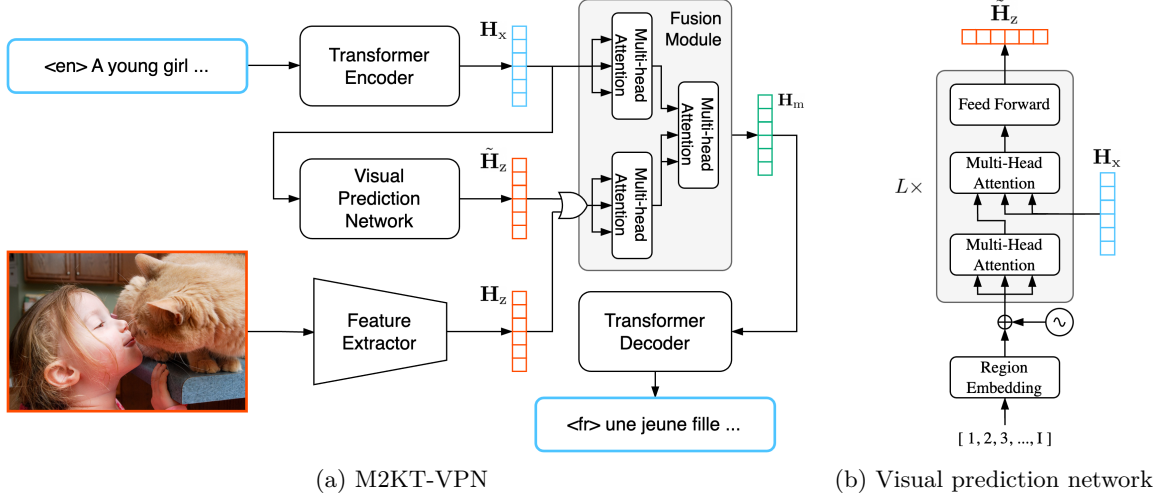
(a) M2KT-VPN   (b) Visual prediction network

Figure 2: The overview of the M2KT-VPN model (a) and the visual prediction network (b).

3.1 Multilingual Machine Translation Module

We design M2KT-VPN as a multilingual MMT model. Following Fan et al. (2021), we prepend a special token (*e.g.*, `<en>`) to the source sentence **x** indicating the source language, and another special token (*e.g.*, `<fr>`) to the target sentence **y** indicating the target language. Similarly, for inference, we condition the decoder to generate a translation in a given target language by prepending its language indicator token as the first token of the sequence to be generated. We employ a cross-entropy loss to train M2KT-VPN models.

3.2 Attention-based Fusion Module

The Transformer encoder embeds a source text **x** into a high-dimensional representation $\mathbf{H}_x \in \mathbb{R}^{N \times d}$ without any presence of images. We then introduce a fusion module to ground the text-only representation $\mathbf{H}_x$ into the image z through its corresponding visual feature $\mathbf{H}_z$. This grounded representation of the source sequence $\mathbf{H}_m \in \mathbb{R}^{N \times d}$ constitutes the input to the Transformer decoder.

We use an attention-based module to fuse the visual input into multimodal representations of language. Our module first applies two dedicated self-attention operations on the text and visual features:

$$\mathbf{H}'_x = \mathrm{MHA}(\mathbf{H}_x, \mathbf{H}_x, \mathbf{H}_x) \quad (1)$$
$$\mathbf{H}'_z = \mathrm{MHA}(\mathbf{H}_z, \mathbf{H}_z, \mathbf{H}_z) \quad (2)$$

where MHA denotes the multi-head attention

function (Vaswani et al., 2017). Then, a cross-attention module fuses these representations to get the multimodal representation $\mathbf{H}_m$:

$$\mathbf{H}_m = \mathrm{MHA}(\mathbf{H}'_x, \mathbf{H}'_z, \mathbf{H}'_z) \quad (3)$$

3.3 Visual Prediction Network

As described so far, our MMT model assumes the input is complete, having both text and image available for translation, both during training and inference. However, in the zero-shot cross-modal machine translation task, the visual modality is absent during training for the zero-shot language pairs.

To mitigate this gap, we propose a **Visual Prediction Network** (**VPN**) to mimic visual features for zero-shot language pairs during training. The VPN generates visual predictions from the text encoder representation $\mathbf{H}_x$. The generated visual predictions $\tilde{\mathbf{H}}_z$ in a zero-shot pair are then fed into the fusion module instead of the visual feature $\mathbf{H}_z$.

To predict the visual features corresponding to I image regions, VPN first embeds learnable visual queries (*e.g.*, Lee et al., 2018; Alayrac et al., 2022; Mañas et al., 2023; Li et al., 2023), adds positional information, and then applies layer normalization to obtain the position-aware region representations $\tilde{\mathbf{H}}^0_z$.

$$\tilde{\mathbf{H}}^0_{z,i} = \mathrm{LayerNorm}(\mathbf{E}_z(i) + \mathrm{PE}(i)) \quad (4)$$

where $\mathbf{E}_z(i)$ is the embedding representation for the i-th region, and $\mathrm{PE}(i)$ is the positional embedding for the i-th region.

The following L layers are the same as in a standard Transformer decoder, each comprising a self-attention, cross-attention, and a pairwise feed-forward module.[1] The l-th layer takes the output of the previous layer $\tilde{\mathbf{H}}_z^{l-1}$ as input. The cross-attention module in the l-th layer takes the output of the self-attention module as the query and the text encoder output $\mathbf{H}_x$ as the key and value. The M2KT-VPN model uses the output representation of the final layer as the visual prediction:

$$\tilde{\mathbf{H}}_z = \tilde{\mathbf{H}}_z^L \tag{5}$$

The VPN module is trained on grounded language pairs, using a max-margin loss (Elliott and Kádár, 2017) in a contrastive learning manner (Radford et al., 2021a). Given a batch of K examples, we first generate K $(\tilde{\mathbf{H}}_z, \mathbf{H}_z)$ pairs. We then compute a max-margin loss for the batch:

$$\sum_{p \neq k}^{K} \sum_{i=1}^{I} \max\{0, \alpha - d(_k\tilde{\mathbf{H}}_{z,i}, _k\mathbf{H}_{z,i}) \\ + d(_k\tilde{\mathbf{H}}_{z,i}, _p\mathbf{H}_{z,i})\} \tag{6}$$

where $_j\tilde{\mathbf{H}}_{z,i}$, $_j\mathbf{H}_{z,i}$ is the predicted i-th vector and the true i-th vector of j-th example in the batch; d is a cosine similarity function; and α is the margin[2]. The max-margin loss is merged with the cross-entropy loss with a coefficient of 1.0 to obtain the final loss.

4 Experiments

4.1 Experimental Setting

Dataset. We train and evaluate models on `Multi30K` dataset. We select English–Czech as a grounded language pair and English–French as a zero-shot language pair. For the training, we divide the training split of `Multi30K` into two folds of the same size; one for the grounded language pair and the other for the zero-shot language pair. The validation splits for grounded and zero-shot language pairs have the same source language texts and the target language texts, but images are absent for the zero-shot language pair. The test splits are also the same, and images are available for both grounded and zero-shot language pairs. Table

---

[1]We use $L = 1$ in our experiments.
[2]We use $\alpha = 0.1$ in our experiments.

Split	Images	Sents.
Grounded (English–Czech)		
Training	14,500	14,500
Validation	1,014	1,014
Test	2,071	2,071
Zero-shot (English–French)		
Training	–	14,500
Validation	–	1,014
Test	3,532	3,532

Table 2: The number of examples in each split for the grounded and zero-shot language pairs.

2 shows the statistics of each split. We follow a standard evaluation to report performance on four test sets: test_2016_flickr (**2016**), test_2017_flickr (**2017**), test_2017_mscoco (**mscoco**), and test_2018_flickr (**2018**).

Preprocessing. For textual modality, we use Moses (Koehn et al., 2007) to lowercase, normalize punctuation, and tokenize the source and target sentences. We then learn byte pair encoding (Sennrich et al., 2016) with 10,000 merge operations on the concatenation of the training text over all language pairs to obtain a shared vocabulary for all languages. For visual modality, we extract a visual feature using **DETR-ResNet-50-DC5**[3] (Carion et al., 2020), which is an object detection model backed by a **ResNet-50** model (He et al., 2016). **DC5** stands for dilated C5 stage, which increases the feature resolution and consequently provides more information for the small objects. The extracted feature has 100 bounding boxes, each with a visual representation of 256 dimensions.

Model. We use a tiny version of the Transformer model (Transformer-tiny) as our text-only baseline and the relying model of M2KT-VPN, as this smaller model works better on `Multi30K` (Wu et al., 2021; Li et al., 2022b). This model comprises four encoder layers and four decoder layers, and the model hidden size of both decoder and decoder is 128. It also has a smaller number of attention heads and a hidden size of pair-wise feedforward network, 4 and 256, respectively. The vocabulary and

---

[3]facebook/detr-resnet-50-dc5

embedding weights are shared across all languages. We compare our model against some baseline models:

- Transformer: a text-only Transformer-tiny model trained only on English–French data.

- mTransformer: a text-only multilingual Transformer-tiny model trained on both English–Czech and English–French data.

- IMAGINATION: a text-only multilingual Transformer with a VPN module. This model also trained on both English–Czech and English–French data.

Implementation details. We implement our models on the Fairseq (Ott et al., 2019) toolkit. The optimizer is Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate warms up from $1e - 7$ to 0.005 over $2,000$ steps, then decays with the `inverse_sqrt` scheduler. We apply label smoothing of 0.1 for computing the cross-entropy loss and the dropout of 0.3. Early stopping with a patience of 10 is used to stop training models. We average the last ten checkpoints and use beam search with width=5 for inference.

Metrics. We train all models three times with different seeds and report averaged 4-gram BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) scores for all test sets. Additional to the classic n-gram matching evaluation, we also compute the COMET score (Rei et al., 2020)[4]. We also report statistical significance ($p < 0.05$) on the difference in BLEU scores[5].

4.2 Results

The results of our experiments are shown in Table 3. We found that our M2KT-VPN model provides an improvement over the text-only baselines and IMAGINATION model for all four test sets. The M2KT-VPN model achieves an averaged improvement of 2.65% over the mTransformer model (varies from 1.90% to 3.70% across the test sets). This performance

gain would be owed to the multitask learning of the visual prediction network; the module learns to predict visual features and tailor the features for the machine translation task simultaneously.

5 Discussion

This section first provides two basic analyses of the M2KT-VPN model: model analysis and probing. We then examine various kinds of features to investigate the importance of feature selection. Finally, we ran an analysis to identify the requirement for the grounded language pair.

5.1 Model Analysis

Model ablation. Table 4 shows the results of a comprehensive ablation analysis to identify the contribution of each module in the M2KT-VPN model on entire test splits. To evaluate the contribution of the attention-based fusion module, we compare two well-known fusion strategies: concatenation-based (Li et al., 2021) and gate-based (Li et al., 2021). Firstly, the model without a VPN module drops -1.0 METEOR score, indicating a VPN module is key to resolving the missing visual modality problem in the zero-shot cross-modal machine translation task. Second, concatenation-based and gate-based models do not outperform the M2KT-VPN model and even the mTransformer baseline. The concatenation-based model fails to translate most of the examples. This evidences that attention-based fusion strategies indeed transfer multimodal knowledge.

Quality of visual prediction. Another question on M2KT-VPN is whether the visual prediction network can provide grounded visual features. To answer this question, We measured each model's Median rank score (Elliott and Kádár, 2017) on the **2016** test data. We first average true and predicted features over their regions to get every single representative vector. The predicted representative vector is compared against the true representative vectors in the test data using the cosine similarity function to produce a ranked order of the true representative vectors. The Median Rank score reports the median value of the ranks for the gold representative vector compared to the predicted representative vector.

---

[4]We use **Unbabel/wmt22-comet-da** (Rei et al., 2022).

[5]We used Moses' **bootstrap-hypothesis-differencesignificance.pl.**

Model	2016	2017	mscoco	2018	Average
Transformer	55.77 / 76.91	47.48 / 70.77	38.95 / 64.12	33.11 / 60.37	43.83 / 68.04
mTransformer	56.42 / 77.57	48.54 / 72.31	40.50 / 65.56	34.31 / 61.67	44.94 / 69.28
IMAGINATION	57.11 / 77.85	49.53 / 72.68	40.75 / 65.95	35.12 / 62.84	45.63 / 69.83
M2KT-VPN	**57.49 / 78.15**	†**50.19 / 73.43**	†**41.28 / 66.44**	†**35.58 / 63.06**	†**46.13 / 70.27**

Table 3: The BLEU / METEOR scores of the text-only models and MMT models in each test set for English–French translation using English–Czech as the grounded language pair. "†" indicates statistical significance of the improvement over the IMAGINATION model.

Fusion Module	VPN	BLEU	METEOR
Attention		44.79	69.27
Concatenation	✓	6.43	19.29
Gate	✓	44.88	69.42
Attention	✓	46.13	70.27

Table 4: The average BLEU and METEOR scores over all test splits for variants of M2KT-VPN.

Model	Median Rank
IMAGINATION	45.5
M2KT-VPN	47.0
Elliott and Kádár (2017)	11.0
Random	~ 500

Table 5: Median rank of randomly selected vector (Random) and model's predictions.

Model	COMET
Transformer	0.7629
mTransformer	0.7651
IMAGINATION	0.7679
M2KT-VPN	0.7698

Table 6: The averaged COMET scores over all test splits for the English–French translation.

Model	2016	2018	Average
Transformer	55.85	47.54	51.69
mTransformer	57.01	49.85	53.43
IMAGINATION	57.99	50.14	54.07
M2KT-VPN	57.78	50.79	54.28

Table 7: The METEOR scores of the text-only and MMT models in each test set for English–Czech translation

Our M2KT-VPN model returns a median rank of 47.0, which is clearly better than the random baseline. This indicates that our model is learning visually grounded representations. However, Elliott and Kádár (2017) reported a median rank of 11.0 for their RNN-based model that predicts holistic features. This difference poses another challenge to predicting region-based visual features using VPN. We would like to improve the prediction quality and explore its impact on the translation quality in our future work.

Neural-based evaluation. Table 6 shows the average COMET score over all test splits. We can see the same trend as BLEU and ME-TEOR in Table 3. While neural-based evaluation metrics would better align with human preference than those based only on surface characteristics, this pattern may vary (Freitag et al., 2021). A human evaluation may rather be conducted to reveal which metrics align bet-

ter with the text of captions, where the text is usually shorter and simpler than those in the WMT evaluation task.

Multilingualism. The multilingualism of the M2KT-VPN model is another concern. Table 7 shows the METEOR score for the English–Czech translation. The consistent improvement over the text-only baseline for both English–Czech and English–French indicates that the M2KT-VPN model is capable of performing multilingual translation.

5.2 Probing

Input degradation. We examine the model's capability of handling incomplete textual modality. Intuitively, a better MMT model can recover the content in the flawed source text from the visual modality. Following Caglayan et al. (2019) and Li et al. (2022a), we

---

[6]"man", "woman", "people", "mean", "girl", and "boy".

	Vanilla	a	young	girl	standing	⋯	a	yellow	cat
Color	a	young	girl	standing	⋯	a	[v]	cat
Entity	a	young	girl	standing	⋯	a	yellow	[v]
Char.	a	young	[v]	standing	⋯	a	yellow	cat
Prog.	a	young	girl	standing	⋯	[v]	[v]	[v]

Table 8: An example of textual degradation. "Vanilla" shows the original text without degradation. "Char." and "Prog." stand for character and progressive masking, respectively. "Color" deprivation replaces words that refer to colors with a special token [v]. "Entity" and "Char." mask out the visually depictable entities and character words[6], respectively. "Prog." masking all words except the first K words. The tokens at "[v]" are masked during both training and inference.

Model	Vanilla	Color	Entity	Char.
mTransformer	77.57	71.85	61.11	70.73
M2KT-VPN	78.15	72.28	61.49	70.78
	(0.03)	(-0.13)	(-0.32)	(0.04)

Table 9: The METEOR scores on vanilla, color-deprivation, entity-masking, and character-masking test sets. The scores in the parenthesis show the METEOR changes when the MMT model takes random shuffled images as its input.
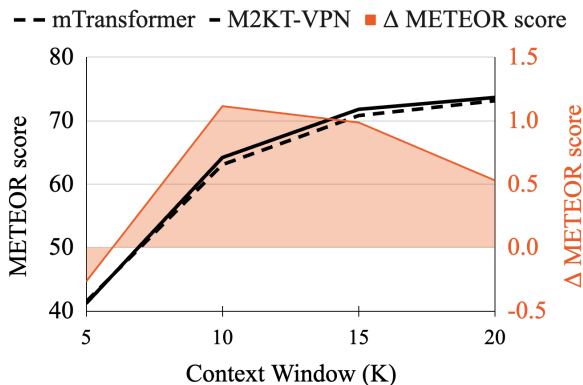

Figure 3: Evaluation with progressive masking of the context size of $\{5, 10, 15, 20\}$.

conducted four kinds of textual degradations: color deprivation, entity masking, character masking, and progressive masking. Table 8 shows examples of a complete text ("Vanilla") and its degraded ones. As entity masking is available only for the **2016** test set, we report all scores only for **2016** test set. Both the training and the test data are degraded.

Table 9 shows the BLEU and METEOR scores of the mTransformer baseline and M2KT-VPN model for vanilla, color-deprived, entity-masked, and character-masked **2016** test sets. The M2KT-VPN model outperforms the mTransformer baseline for color and entity

degradation scenarios, while we see almost no change for character degradation. The possible cause of this difference is the nature of the **DETR** model we used to extract the feature. As the labels that **DETR** learns to predict contain only one word ("person") to stand for characters but more words for entities, an MMT model incorporating **DETR** would be capable of recognizing entities more precisely rather than characters. Table 10 also supports this idea. While the sentence's third [v] (corresponding to "bench") is correctly translated into "vif", the first masked entity (corresponding to "woman") keeps being mistranslated. As shown in the image, the **DETR** feature provides useful information to distinguish the "bench" from the "chair". However, it is not informative to identify the gender of the person in the image.[7]

Figure 3 compares the METEOR scores of the mTransformer baseline and an M2KT-VPN model for progressive-masked **2016** test sets with different context windows (K). The MMT model outperforms the baseline for $K = \{10, 15, 20\}$. The gap between the baseline and MMT model widens at $K = \{10, 15\}$ and narrows at $K = \{5, 20\}$. This observation for $K = \{10, 15, 20\}$ is consistent with a previous work of Li et al. (2022b), which claims the gap widens as the context window is reduced, while that for $K = 5$ is contrary to the claim. This suggests that the visual prediction network could fail to provide rich visual information when the textual context is extremely limited.

Visual awareness. We also examine the reliance of the model on the visual modality. To

---

[7]We found all three trained text-only systems failed to translate [v] corresponding to "bench", and all M2KT-VPN models successfully translate it.
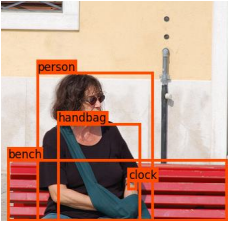
	Vanilla Entity References	the woman in the brown shirt is sitting on a bright red bench . the [v] in the brown [v] is sitting on a bright red [v] . la femme en t-shirt marron est assise sur un banc rouge vif .
mTransformer		l'homme en t-shirt marron est assis sur une chaise de couleur vive . (the man in the brown t-shirt is sitting on a brightly colored chair.)
M2KT-VPN		l'homme en t-shirt marron est assis sur un banc rouge vif . (the man in the brown t-shirt is sitting on a bright red bench.)

Table 10: Translation examples of the baseline and MMT model. The bounding boxes in the image are the prediction of the **DETR-ResNet-50-dc5** model and have a score of above 0.8. We use DeepL to translate each hypothesis into English and show it in each parenthesis.

this end, we compute the performance deterioration when a model receives incongruent images instead of congruent images (Elliott, 2018). The scores with parenthesis in Table 9 show the performance changes when the model takes incongruent images. Without surprise, the MMT model is not aware of images for vanilla, color-deprived, and character-masked test sets, as the **DETR** model does not provide rich information about color and character in an image. Meanwhile, the model is sensitive to the input image when the entities in the source text are masked out; the MMT model readily uses **DETR** feature to disambiguate the masked entities.

5.3 Visual Feature Selection

Selecting a proper visual feature has been proven to affect MMT model performance (Li et al., 2021).

In Table 11, we compare the M2KT-VPN models using different visual features extracted by different vision backbones.

- **ResNet** (He et al., 2016): An image recognition model trained to classify an image into one of the 1,000 ImageNet classes. **ResNet-50** and **ResNet-101** comprise 50 and 101 layers, respectively. We extract the local features of each ResNet model and feed them into the MMT models.

- **Faster R-CNN** (Anderson et al., 2018): An object detection model trained to segment an image into 36 salient image regions and predict the object in each region.

- **DETR** (Carion et al., 2020): A transformer-based object detection model trained to segment an image into 100 regions and predict the object in each region. We used four different backbones:

ResNet-50, **ResNet-50-DC5**, **ResNet-101**, and **ResNet-101-DC5**.

- **CLIP** (Radford et al., 2021b): A vision and language model trained on various image and text pairs in a self-supervised way. We examined three CLIP models using different backbones: **ResNet-101**, **ViT-B/16**, and **ViT-B/32**. We use the visual encoder of each CLIP model to encode images; no textual modality is involved in the extraction process.

10 out of 11 MMT models outperform the mTransformer model in both BLEU and METEOR scores. This shows that M2KT-VPN models are capable of incorporating various kinds of visual features. The only feature that deteriorates the model performance is ResNet-101; the feature extracted by ResNet-101 would be highly optimized for image classification and not suitable for machine translation.

Among all features, **DETR** with the **ResNet-50-DC5** backbone serves as the best feature extractor for the M2KT-VPN model. On the other hand, the model using **CLIP** features obtains almost equal performance to those using **ResNet** features. This observation is partially contrary to the previous works claiming that enhanced vision features obtain superior performance compared with low-level vision features (Li et al., 2022a).

We also observed that **DETR** with **DC5** backbone outperforms the non-DC5 counterparts. As **DC5** models provide the feature with higher resolution, the MMT model can receive richer information about small objects in an image. Consequently, the MMT model can better understand and translate those small objects more accurately.

Feature	BLEU	METEOR
None (mTransformer)	44.94	69.28
ResNet-50	45.34	69.67
ResNet-101	44.79	69.29
Faster R-CNN	45.72	69.65
DETR (ResNet-50)	45.79	70.01
DETR (ResNet-50-DC5)	**46.13**	**70.27**
DETR (ResNet-101)	45.49	69.84
DETR (ResNet-101-DC5)	45.81	69.91
CLIP (ResNet-101)	45.19	69.47
CLIP (ViT-B/16)	45.64	69.88
CLIP (ViT-B/32)	45.36	69.64

Table 11: The averaged BLEU and METEOR scores over all test splits of M2KT-VPN models using different visual features. The models in the parentheses are backbone models.

Grounded	BLEU	METEOR
→ Czech	46.14 (↑1.12)	70.27 (↑0.93)
→ German	45.95 (↓2.04)	69.95 (↓1.16)
→ Japanese	42.34 (↓2.53)	68.25 (↓0.99)

Table 12: The scores over all test splits of the M2KT-VPN model using different grounded language pairs. Each "→ X" stands for English → X as the grounded language pair. The scores in parenthesis are the changes from the text-only counterpart.

5.4 Grounded Language Pairs

The ability of a model to transfer multimodal knowledge between grounded and zero-shot language pairs is another key research question for this task. To answer this question, we compare three grounded language pairs for English–French zero-shot cross-modal translation.[8]

Shown in Table 12, the translation performance of using English–Czech as a grounded language pair is better than those of using English–German and English–Japanese.

The observation of using English–German contradicts our intuition that the more similar two language pairs are, the better one serves as a grounded language pair for another. As English–German training data is generated with no involvement of images, this indicates that M2KT-VPN requires image-aware training data to transfer multimodal knowledge.

English–Japanese also contains visual-aware translations, but it does not improve the performance of English–French. We found that M2KT-VPN translated the 1.43% of entire test examples into Japanese regardless the decoder is conditioned to generate French translation[9]. This ratio is much higher than that of the text-only counterpart (0.27%) and M2KT-VPN using English–Czech (0.26%) or English–German (0.28%). We conclude that grounded and zero-shot pairs should not be too distant.

6 Related Work

Multimodal machine translation. This task has been developed along with the creation of multimodal parallel corpora. After the first multimodal parallel corpus, namely Multi30K for English–German translation, emerged at the first conference of machine translation (Bojar et al., 2016), many publicly available datasets have been proposed: the English–French version of Multi30K and new test sets at 2017 (Elliott et al., 2017), the English–Czech version of Multi30k (Barrault et al., 2018), and the English–Japanese version of Multi30k (Nakayama et al., 2020). More recently, Guo et al. (2022) proposed a private expansion of Multi30K, including Hindi, Turkish, and Latvian translations. They examined a multilingual MMT model on their dataset and investigated the multilingual ability of the model. We put the step forward and investigate the zero-shot cross-modal translation capability in an MMT task.

Predicting a visual feature from textual modality is a well-established approach for improving multimodal machine translation systems. Elliott and Kádár (2017) first divided the multimodal machine translation task into two subtasks: translation task and visual grounding task. Similarly, Zhou et al. (2018) employed a latent space learning task as their visual grounding task to bridge textual and visual modalities. Recently, Li et al. (2022b) proposed to utilize the feature prediction from a visual prediction network. We make use of the model for the visually grounding task and propose to incorporate the prediction as a pseudo-visual feature with MMT models.

[8]We retrieved Japanese translations from Flickr30kEnt-JP (Nakayama et al., 2020)

[9]We used Google's language-detection library.

Zero-shot cross-lingual machine translation. Zero-shot *cross-lingual* machine translation aims to perform a translation with zero-resource where the considering language pairs do not have any parallel corpora (Firat et al., 2016; Johnson et al., 2017; Chen et al., 2017; Lample et al., 2018; Artetxe et al., 2019). The previous works have proved the zero-shot cross-lingual translation capability.

In a multimodal setting, we are only aware of two previous efforts on zero-shot transfer. Huang et al. (2020) simulated that no parallel corpus exists between the language pair and proposed utilizing the image as the pivot and performing a zero-shot cross-lingual translation. Besides, Long et al. (2021) trained a generative adversarial network (GAN) (Goodfellow et al., 2014) for generating the visual features for text-only language pairs. Both approaches use images for training, and evaluate models on a single text-only translation direction. Unlike these works, our work (i) tests MMT models with complete multimodal inputs and (ii) takes advantage of a multilingual model.

7 Conclusion

In this paper, we proposed a new task, **zero-shot cross-modal machine translation**, aiming to evaluate MMT systems from the perspective of the cross-lingual transferability of multimodal knowledge learned from grounded language pairs into language pairs with only text data during training.

Our proposed MMT model shows promising results, suggesting that the VPN mitigates the modality mismatch between training and inference steps for zero-shot language pairs. The analysis shows the importance of selecting a proper visual feature and the necessity of image-aware translations, both of which should be key properties of MMT models.

Limitations

Although our M2KT-VPN model has shown the zero-shot cross-modal translation capability, some limitations exist. While the well-established visual features are informative for some object entities, they do not benefit the translation of character and color words. Besides, the importance of language similarity between grounded and zero-shot pairs limits the language pairs we can apply M2KT-VPN for. In future work, we will extend our M2KT-VPN model to relax this limitation.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. 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.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg. Springer-Verlag.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, 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*, 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multilingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. 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 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. 2018. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

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.

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

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2523–2548, Dubrovnik, Croatia. Association for Computational Linguistics.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual

models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. WAT2019: English-Hindi translation on Hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

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

X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.

A   Translation Examples

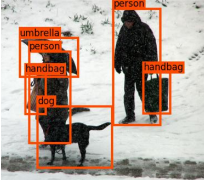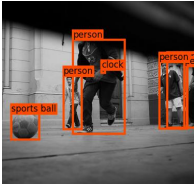Table 13 shows the translation examples for the vanilla source text.

	Source Reference	two people are walking the dog through the snow . deux personnes promènent leur chien dans la neige .
	mTransformer	deux personnes marchent ϕ dans la neige . (two people walking ϕ in the snow .)
	M2KT-VPN	deux personnes promènent le chien dans la neige . (two people walking the dog in the snow .)
	Source Reference	several children are watching someone chase a ball on the sidewalk . plusieurs enfants regardent quelqu'un courir après une balle sur le trottoir .
	mTransformer	plusieurs enfants regardent quelqu'un ϕ sur le trottoir . (several children look at someone ϕ on the sidewalk .)
	M2KT-VPN	plusieurs enfants regardent quelqu'un après une balle sur le trottoir . (several children look at someone after a ball on the sidewalk .)

Table 13: Translation examples of the baseline and M2KT-VPN model for the vanilla source text. The bounding boxes in the image are the prediction of the **DETR-ResNet-50-dc5** model and have a score of above 0.8. We use DeepL to translate each hypothesis into English and show it in each parenthesis. The "ϕ" stands for the omitted target word in the translation.