Soul-Mix: Enhancing Multimodal Machine Translation with Manifold Mixup

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An
Hongxiang Li, Yaowei Li, Yuexian Zou*
School of ECE, Peking University, China
{chengxx, yaozy, xinyifei, anhao, lihongxiang, ywl}@stu.pku.edu.cn
zouyx@pku.edu.cn

Abstract

Multimodal machine translation (MMT) aims to improve the performance of machine translation with the help of visual information, which has received widespread attention recently. It has been verified that visual information brings greater performance gains when the textual information is limited. However, most previous works ignore to take advantage of the complete textual inputs and the limited textual inputs at the same time, which limits the overall performance. To solve this issue, we propose a mixup method termed Soul-Mix to enhance MMT by using visual information more effectively. We mix the predicted translations of complete textual input and the limited textual inputs. Experimental results on the Multi30K dataset of three translation directions show that our Soul-Mix significantly outperforms existing approaches and achieves new state-of-the-art performance with fewer parameters than some previous models. Besides, the strength of Soul-Mix is more obvious on more challenging MSCOCO dataset which includes more out-of-domain instances with lots of ambiguous verbs.

1 Introduction

Multimodal machine translation (MMT) endeavors to enhance the translation systems via integrating different multimodal datasets, specifically from the visual inputs, into conventional text-only neural machine translation (NMT) systems (Caglayan et al., 2019; Yao and Wan, 2020; Yang et al., 2020; Fei et al., 2023; Tayir et al., 2024; Tayir and Li, 2024). When visual context is incorporated into the translation process, the accuracy of the translation is expected to improve significantly. This is because visual context can clarify words that have multiple meanings, which ensures that the intended meaning is conveyed more precisely and accurately.

Most current MMT frameworks aim to address the semantic gap between images and texts by designing fusion frameworks. Caglayan et al. (2021) applies the pre-training technique and proposes a cross-lingual method to learn the visually-grounded representations for MMT. Li et al. (2022a) applies a selective attention framework to correlate images to words. Li et al. (2022b) enhances MMT by introducing visual hallucination during inference time, as opposed to other MMT approaches based on the sentence-image pairs. Huang et al. (2023) contrasts multi-modal input with adversarial samples, in this case, the model learns to identify the most informative sample which is combined with the consistent image and several visual objects extracted from it.

Despite the promising progress achieved by existing MMT methods, most of them do not take full advantage of both complete and limited textual inputs. While recent research has shown that limited textual context can lead to better translations based on visual inputs, most attempts to incorporate limited textual inputs ignore the potential benefits of complete textual inputs (Caglayan et al., 2019). Intuitively, predicted translations based on complete textual inputs tend to rely more on textual information, ensuring semantic accuracy, while predicted translations based on the limited textual inputs tend to incorporate more information from visual inputs, clarifying words with multiple meanings. However, the lack of complete textual inputs could limit the performance when utilizing limited textual inputs.

To tackle this problem, we propose a framework termed Soul-Mix based on manifold mixup (Zhang et al., 2018; Verma et al., 2019) to enhance MMT by leveraging visual information more effectively. Specifically speaking, we mix the predicted translations of the complete textual input and the limited textual inputs. The mixed translation and the translation of the complete textual input are both utilized to compute the cross-entropy loss. In addition, we utilize Jensen-Shannon Divergence (JSD) to regu-

* Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11283–11294

larize their predictions to further boost the performance. Experiment results on the Multi30K (Elliott et al., 2016) dataset of English-to-German, English-to-French, and English-to-Czech translation directions demonstrate that our framework outperforms previous MMT methods, particularly on the more challenging MSCOCO dataset (Elliott et al., 2017), which includes more out-of-domain instances and ambiguous verbs. Further model analysis also verifies the advantage of our approach.

To sum up, our contributions are three-fold:

- We propose a manifold mixup method to use visual information more effectively.
- Experiment results show that our framework achieves the new state-of-the-art performance with only a small increase in parameters.
- Further model analysis demonstrates that our framework is more robust to noise from irrelevant visual input information.

2 Related Work

Related work primarily includes the studies on multimodal machine translation and manifold mixup.

2.1 Multimodal Machine Translation

Machine translation aims to convert text or speech from one language to another (Chen et al., 2022; Cheng et al., 2023b; Chen et al., 2024; Gui et al., 2024). As a language shared by people worldwide, the visual modality helps machines to have a more comprehensive perception of the real world. The superiority of the multimodal context in MMT has been illustrated in recent works. Therefore, there has been a growing interest in MMT tasks (Elliott et al., 2016), which attracts more and more attention. As pre-trained models have great potential in other tasks (Cao et al., 2022; Li et al., 2023a,b; Jin et al., 2023; Dong et al., 2023a; Cai et al., 2023; Zhu et al., 2023b; Dong et al., 2023b; Cai et al., 2024; Huang et al., 2024), pre-trained models are also widely leveraged in MMT, such as ResNet (He et al., 2016). To prevent the encoding of irrelevant information in images, Yao and Wan (2020) proposes multimodal self-attention in the Transformer, which allows the model to effectively capture the relevant information in these modalities. Nishihara et al. (2020) introduces the supervised visual attention mechanism that leverages manual alignments between words in an utterance and their corresponding regions in the given image. This approach could generate supervisions for visual attention and trains the visual attention component of the encoder utilizing these supervisions, boosting the model's ability to align textual and visual information accurately. Huang et al. (2020) focuses on the huge challenge of unsupervised MMT by incorporating pseudo visual pivoting and the visual inputs. Wu et al. (2021) proposes two innovative approaches for MMT. The first method, gated fusion, is designed for conventional MMT, which fuses visual and textual inputs. The second method, dense-retrieval augmentation, leverages the retrieval-based approach, incorporating dense retrieval mechanisms to boost translation accuracy via retrieving the relevant visual and textual data. Li et al. (2022a) develops a new selective attention mechanism that correlates image patches with words, which has been widely utilized in recent research, as it allows for more precise alignment between visual and textual elements, thereby improving the overall performance of multi-modal translation systems. Ye and Guo (2022) leverages a mixup strategy to extract useful visual features, enhancing the machine translation process. This technique improves the model's ability to use the visual information effectively, leading to better translation outcomes. Ye et al. (2022) designs the novel robust multi-modal interactive fusion method, including the cross-modal relation-aware mask mechanism. It enhances the model's robustness to noise and improves the interaction between different modalities, leading to more accurate and more reliable translations. Cheng et al. (2023d) leverages asymmetric contrastive learning at these two levels to leverage the knowledge from image captioning and object detection. In this paper, we propose to utilize manifold mixup (Zhang et al., 2018; Verma et al., 2019) to transfer knowledge from visual information.

2.2 Mixup

How to improve the model's robustness is a topic worth discussing (Xin et al., 2022; Cheng et al., 2023a; Yin et al., 2023; Cheng et al., 2023c; Xin et al., 2023b; Xin and Zou, 2023; Xin et al., 2023a; Zhu et al., 2024; Zhuang et al., 2024). Zhang et al. (2018) first proposes mixup to boost the model's robustness. Beyond this, Verma et al. (2019) builds manifold mixup via applying this technique to the hidden representations, extending the surface-level mixup method. Some recent works have introduced mixup on machine translation (Zhang et al., 2019; Li et al., 2021), sentence classification (Chen et al., 2020; Jindal et al., 2020; Sun et al., 2020), multilingual understanding (Yang et al., 2022), speech

SRC:	a	man	in	a	red	shirt	entering	an	establishment
Color	a	man	in	a	[MASK_C]	shirt	entering	an	establishment
Character	a	[MASK_P]	in	a	red	shirt	entering	an	establishment
Noun	a	man	in	a	red	shirt	entering	an	[MASK_N]

Figure 1: An example of the degradation schemes for textual inputs. Three special tokens are designed to replace the specific words in the given sentence respectively.

recognition (Medennikov et al., 2018; Meng et al., 2021), and sentiment analysis (Zhu et al., 2023a). Note that Ye and Guo (2022) also applies mixup in MMT. However, we focus on mixing the representations of the complete textual inputs and limited textual inputs to enhance MMT.

3 Method

In this section, we first introduce the problem defination (§3.1) and the degradation schemes (§3.2). Then, we introduce our proposed Soul-Mix, including the source sentence encoder (§3.3), the aggregation module (§3.4), the target sentence decoder (§3.5), and the manifold mixup module (§3.6). At last, we introduce the final training objective (§3.7). The overview of our approach is shown in Figure 2.

3.1 Problem Defination

The corpus of MMT usually contains triples, which could be denoted as $\mathcal{D} = \{(x, y, z)\}$, where z denotes the image and x, y denote the corresponding description of the image in the source language and target language. Given $\{(x, z)\}$, the cross-entropy loss is formulated as follows:

$$\mathcal{L}_{\text{CE}}(x, y, z) = -\sum_{i=1}^{|y|} \log p(y_i | y_{<i}, x, z) \quad (1)$$

3.2 Degradation Schemes

As shown in Figure 1, when dealing with textual inputs, we adopt the degradation schemes (Caglayan et al., 2019; Li et al., 2022a), which involves using three degradation schemes: color masking, character masking, and noun masking.

Color masking refers to the process of substituting all the words that denote a specific color with a special token `[Mask_C]`, which aims to boost the ability to capture color information from vision.

Character masking means that we replace character words by a special token `[Mask_P]`. Following Li et al. (2022a), we regard "man", "woman", "people", "men", "girl", and "boy" as the character words, which aim to enhance the ability to capture gender information from vision. This is because, in some languages, the masculine and feminine forms of the same noun are different. For example, "Eine Baseballspielerin" in German means *a female baseball player*, while "Ein Baseballspieler" in German means *a male baseball player*.

Noun masking means that all nouns contained in the utterance might be replaced. However, since noun masking is a more complex scenario than the character masking, we only mask one of the nouns using a special token `[Mask_N]`.

It is worth noting that while the majority of sentences in the Multi30K dataset contain a color word, a character word, and a noun, there are still some utterances that do not include any of these words. In such cases, we directly use the original sentence as the output after applying the degradation process.

3.3 Source Sentence Encoder

As illustrated in the left part of Figure 2, our source sentence encoder follows the same structure as the transformer encoder (Vaswani et al., 2017). Given the source sentence $x = \{x_1, \cdots, x_n\}$, the output of the source sentence encoder is as follows:

$$\mathrm{E}_x^l = \mathrm{FFN}(\mathrm{Multihead}(\mathrm{E}_x^{l-1}, \mathrm{E}_x^{l-1}, \mathrm{E}_x^{l-1})) \quad (2)$$

where E_x^0 denotes the embedding with the position embedding, n denotes the length of x, and l denotes the numbers of Encoder layers.

3.4 Aggregation Module

As illustrated in the central part of Figure 2, following Zhang et al. (2020); Yin et al. (2020); Cheng et al. (2023d) we introduce an aggregation module

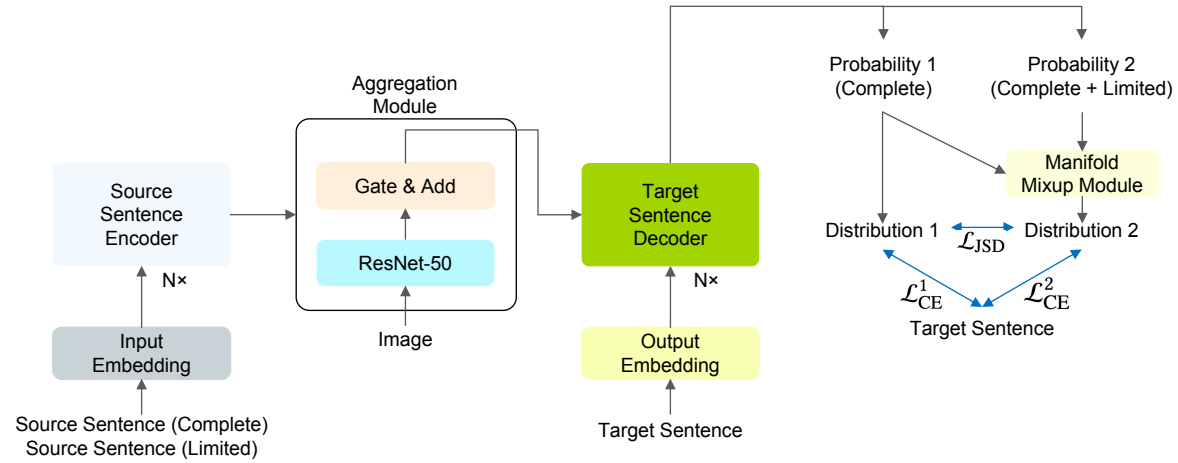

Figure 2: Overview of our framework. The input of probability 1 includes the complete source sentence and the image. The input of probability 2 includes the complete source sentence, the limited source sentence, and the image.

to fuse these two modalities. For the input image z, a pre-trained ResNet-50 CNN (He et al., 2016) is utilized to obtain the average-pooled visual representation, which is projected to the same dimension as E_x. The projected visual representation E_z is:

$$\mathrm{E}_z = \mathbf{W}_z \, \mathrm{ResNet}_{\mathrm{pool}}(z). \quad (3)$$

where $\mathbf{W}_z$ is the parameter matrix. A gating matrix Λ is introduced to control the fusion of E_x and E_z:

$$\Lambda = \mathrm{sigmoid}(\mathbf{W}^1_\Lambda \mathrm{E}_x + \mathbf{W}^2_\Lambda \mathrm{E}_z) \quad (4)$$

where $\mathbf{W}^1_\Lambda$ and $\mathbf{W}^2_\Lambda$ are two parameter matrices.

After fusion, we obtain the output E as follows:

$$\mathrm{E} = \mathrm{E}_x + \Lambda \mathrm{E}_z \quad (5)$$

E is fed into the target sentence decoder.

3.5 Target Sentence Decoder

As illustrated in the right part of Figure 2, our target sentence decoder follows the same structure as the one proposed in Vaswani et al. (2017). For the output of the aggregation module E and the embedding of the input target sentence D^0, the output of each layer D^l in the target sentence decoder is:

$$\mathrm{H}^l_1 = \mathrm{Multihead}(\mathrm{D}^{l-1}, \mathrm{D}^{l-1}, \mathrm{D}^{l-1}) \quad (6)$$

$$\mathrm{H}^l_2 = \mathrm{Multihead}(\mathrm{H}^l_1, \mathrm{E}, \mathrm{E}) \quad (7)$$

$$\mathrm{D}^l = \mathrm{FFN}(\mathrm{H}^l_2) \quad (8)$$

where H^l_1 is the output of the l-th self-attention of the decoder, H^l_2 is the output of the cross attention layer in the l-th layer, and D^l is the hidden states in the l-th layer of the decoder. Then a softmax layer is applied to generate the probability distribution, which takes the hidden states D in the top layer of the target sentence decoder as the input.

3.6 Manifold Mixup Module

Previous studies have demonstrated that when the textual input is limited, MMT models could leverage visual information more effectively (Caglayan et al., 2019). In this work, we extend this method by utilizing a manifold mixup technique to simultaneously leverage the predicted translations of the limited textual inputs and the predicted translation of the complete textual input. Specifically, we apply several degradation schemes to textual inputs as Section 3.2. For the complete input sentence x^0, we utilize color masking, character masking, and noun masking, to obtain the processed sentences x^1, x^2, and x^3, respectively.

We utilize $\{(x^i, z^0)\}^3_{i=0}$ as the input of the encoder and the pre-trained ResNet-50 CNN, respectively, and apply manifold mixup technique to mix the predicted translations of each pair of inputs. The mixed predicted translation y^{mix} is as follows:

$$y^{\mathrm{mix}} = \sum_{i=1}^{3} \lambda \mathbf{F}(x^i, z^0) + (1 - 3\lambda)\mathbf{F}(x^0, z^0) \quad (9)$$

where $\mathbf{F}(\cdot)$ denotes the corresponding output of the target sentence decoder and λ is a hyper-parameter of the manifold mixup technique.

3.7 Training Objective

The training objective of our framework includes two cross-entropy losses, where the first is calculated between the predicted translation y^{com} of complete textual input and the ground truth y, and the second is calculated between the mixed predicted translation y^{mix} and the ground truth y, which can

Models	#Params	Test2016		Test2017		MSCOCO	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
English-to-German							
Transformer† (Vaswani et al., 2017)	2.6M†	41.02	68.22	33.36	62.05	29.88	56.64
Imagination† (Elliott and Kádár, 2017)	7.0M†	41.31	68.06	32.89	61.29	29.90	56.57
Multimodal Self-attn‡ (Yao and Wan, 2020)	–	41.50	58.52	32.51	51.33	29.10	48.48
WRA-guided♢ (Zhao et al., 2021)	–	39.30	58.30	32.30	52.80	28.50	48.50
Selective Attn♢ (Li et al., 2022a)	–	41.93	68.55	33.60	61.42	31.14	56.77
DLMulMix♢ (Ye and Guo, 2022)	–	41.77	58.93	33.07	51.85	29.90	49.09
PLUVR♢ (Fang and Feng, 2022)	–	40.30	–	33.45	–	30.28	–
VALHALLA(M)♢ (Li et al., 2022b)	–	42.60	69.30♦	35.10	62.80♦	30.70♦	57.60♦
E2H-MNMT♢ (Ye et al., 2023)	–	42.84♦	60.16	35.60♦	55.00	30.56	50.91
RG-MMT-EDC♢ (Tayir et al., 2024)	–	42.00	60.20	33.40	53.70	30.00	49.60
Soul-Mix (ours)	3.0M	**44.24**♣	**69.93**♣	**37.14**♣	**63.59**♣	**34.26**♣	**59.94**♣
English-to-French							
Transformer† (Vaswani et al., 2017)	2.6M†	61.80	81.02	53.46	75.62	44.52	69.43
Imagination† (Elliott and Kádár, 2017)	6.9M†	61.90	81.20	54.07	76.03	44.81	70.35
Multimodal Self-attn‡ (Yao and Wan, 2020)	–	61.44	75.77	54.56	71.62	44.59	65.08
WRA-guided♢ (Zhao et al., 2021)	–	61.80	76.30	54.10	70.60	43.40	63.80
Selective Attn♢ (Li et al., 2022a)	–	62.48	81.71	54.44	76.46	44.72	71.20
DLMulMix♢ (Ye and Guo, 2022)	–	62.23	76.85	55.18	73.37	44.42	66.41
PLUVR♢ (Fang and Feng, 2022)	–	61.31	–	53.15	–	43.65	–
VALHALLA(M)♢ (Li et al., 2022b)	–	63.10	81.80♦	56.00	77.10♦	46.40	71.30♦
E2H-MNMT♢ (Ye et al., 2023)	–	63.36♦	77.29	56.35♦	72.76	47.04♦	67.36
RG-MMT-EDC♢ (Tayir et al., 2024)	–	62.90	77.20	55.80	72.00	45.10	64.90
Soul-Mix (ours)	3.0M	**64.75**♣	**83.24**♣	**57.47**♣	**78.23**♣	**49.25**♣	**73.48**♣

Table 1: The numbers of parameters, BLEU scores, and METEOR scores on the Multi30k dataset of the English-to-German and the English-to-French translation direction. Results with "♢" denote that they are taken from the corresponding published papers, and results with † are cited from Wu et al. (2021). '–' denotes missing results from the published work. ♦ denotes the previous best results. ♣ denotes our framework significantly outperforms the baselines with $p < 0.01$ under t-test. The best results are highlighted in bold.

be formulated as follows:

$$\mathcal{L}^1_{\text{CE}} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, x^0, z^0) \tag{10}$$

$$\mathcal{L}^2_{\text{CE}} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, x^0, x^1, x^2, x^3, z^0) \tag{11}$$

Besides the cross-entropy losses, we also regularize the above two output predictions y^{com}, y^{mix} by minimizing JSD between the two output distributions, which is formulated as follows:

$$\mathcal{L}_{\text{JSD}} = \sum_{i=1}^{|y|} \text{JSD}\{p(y_i|y_{<i}, x^0, z^0) \| p(y_i|y_{<i}, x^0, x^1, x^2, x^3, z^0)\} \tag{12}$$

The final training objective is as follows:

$$\mathcal{L} = \mathcal{L}^1_{\text{CE}} + \mathcal{L}^2_{\text{CE}} + \gamma \mathcal{L}_{\text{JSD}} \tag{13}$$

where γ is the coefficient weight to control $\mathcal{L}_{\text{JSD}}$.

4 Experiments

4.1 Datasets and Metrics

All experiments are conducted on the Multi30K[1] (Elliott et al., 2016) dataset, which is an extension of Flickr30k (Young et al., 2014) widely utilized for MMT. Each image in Multi30K is accompanied by a sentence in English (En) and the manual translations in German (De), French (Fr), and Czech (Cs). The training set contains 29,000 text-image pairs and the validation set contains 1014 pairs. Four test

[1] https://github.com/multi30k/dataset

Models	Test2016		Test2018	
	BLEU	METEOR	BLEU	METEOR
Transformer† (Vaswani et al., 2017)	32.70	32.34	27.62	29.03
Doubly-ATT† (Arslan et al., 2018)	33.25	32.28	29.12	29.87
Multimodal Self-attn† (Yao and Wan, 2020)	33.12	32.01	28.75	29.51
Gated Fusion† (Wu et al., 2021)	33.77	32.24	29.43	29.41
NR-MNMT♢ (Ye et al., 2022)	35.09	33.52♦	31.40♦	31.26♦
E2H-MNMT♢ (Ye et al., 2023)	35.18♦	33.39	31.29	30.82
Soul-Mix (ours)	**36.45**♣	**34.73**♣	**32.77**♣	**32.65**♣

Table 2: BLEU scores and METEOR scores on the Multi30k dataset of the English-to-Czech translation direction. Results with "♢" denote that they are taken from the corresponding published papers and results with † are cited from Ye et al. (2022). ♦ denotes the previous best results. ♣ denotes our framework significantly outperforms the baselines with $p < 0.01$ under t-test. The best results are highlighted in bold.

sets are used to evaluate the MMT models, including Test2016 (Elliott et al., 2016), Test2017 (Elliott et al., 2017), MSCOCO (Elliott et al., 2017), and Test2018 (Barrault et al., 2018). The details of the dataset can be found in Appendix. A.

To evaluate the performance of translation, we report 4-gram BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). 4-gram BLEU measures the quality of translations in accuracy and fluency and METEOR takes into account both precision and recall. Higher BLEU and METEOR mean higher performance.

4.2 Implementation Details

Following the previous works, the byte pair encoding (BPE) algorithm (Sennrich et al., 2016) with 6,000 merging operations is used to segment words into subwords, which could generate a vocabulary of 5876 tokens for En-De, 5684 tokens for En-Fr, and 5972 tokens for En-Cs. A pre-trained ResNet-50 CNN (He et al., 2016) is utilized to extract the image features. Both the encoder and decoder contain 4 layers. Sign test (Collins et al., 2005) is used as the standard statistical-significance test.

During the training process, the value of label smoothing is set to 0.1, and dropout is set to 0.1. The weight λ is set to 0.2 and the weight γ is set to 2. During decoding, the beam size is set to 5. For all the experiments, we select the model that works the best on the `dev` set and then evaluate it on the `test` set. To avoid overfitting, the training will early-stop if the loss on `dev` set does not decrease for 3 epochs as in (Zhang et al., 2020). We apply Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.98$, and 4k warm-up updates to optimize all the parameters, where we linearly increase the learning rate from 10e-7 to 5e-5.

During the inference process, we follow the previous work (Wu et al., 2019) to average the checkpoints of the last 10 epochs for evaluation. Multi-bleu.perl8 script[2] is used to compute case-sensitive 4-gram BLEU scores for all `test` sets. The entire training process takes several hours. All the experiments are conducted on a single Nvidia V100 GPU with fp16 and based on fairseq[3] (Ott et al., 2019).

4.3 Baselines

We first compare our method with a text-only baseline, which trains a transformer model without any visual information. In addition, we report the performance of several MMT models, including Imagination, Multimodal Self-attn, WRA-guided, Selective Attn, DLMulMix, PLUVR, VALHALLA(M), CAT-MMT bi, E2H-MNMT, and RG-MMT-EDC. For a fair comparison, we utilize the same configuration for all the baselines as our framework.

4.4 Results

Table 1 demonstrates the performance comparison between Soul-Mix and the baselines in English-to-German, and English-to-French translation directions. We also report the result in English-to-Czech translation direction in Table 2. Based on these results, we have the following observations:

(1) Soul-Mix achieves consistent improvements on all the test sets, including the Test2016 dataset, the Test2017 dataset, the Test2018 dataset, and the

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[3] https://github.com/facebookresearch/fairseq

Models	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Soul-Mix	**44.24**	**69.93**	**37.14**	**63.59**	**34.26**	**59.94**
w/o COM	43.79 (↓0.45)	69.70 (↓0.23)	36.81 (↓0.33)	63.40 (↓0.19)	33.12 (↓1.14)	58.96 (↓0.98)
w/o CHM	43.97 (↓0.27)	69.78 (↓0.15)	36.98 (↓0.16)	63.50 (↓0.09)	33.58 (↓0.68)	59.43 (↓0.51)
w/o NM	44.11 (↓0.13)	69.85 (↓0.08)	36.20 (↓0.94)	63.53 (↓0.06)	33.89 (↓0.37)	59.68 (↓0.26)
+ RM	43.81 (↓0.43)	69.67 (↓0.26)	36.87 (↓0.27)	63.38 (↓0.21)	33.22 (↓1.04)	59.18 (↓0.76)

Table 3: Results of different degradation schemes on Multi30k of English-to-German translation direction.

Models	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Soul-Mix	**44.24**	**69.93**	**37.14**	**63.59**	**34.26**	**59.94**
w/o mixup	43.13 (↓1.11)	69.45 (↓0.48)	36.32 (↓0.82)	63.16 (↓0.43)	32.44 (↓1.82)	58.22 (↓1.72)
w/o JSD	43.87 (↓0.37)	69.71 (↓0.22)	37.02 (↓0.12)	63.26 (↓0.33)	33.78 (↓0.48)	59.45 (↓0.49)

Table 4: Results of ablation experiments on the Multi30k dataset of English-to-German translation direction.

MSCOCO dataset. For English-to-German translation direction, our approach achieves 44.24, 37.14, and 34.26 BLEU on these test sets, outperforming the previous results. Similar situations can also be found in the results of METEOR and other translation directions, which demonstrate the effectiveness of our framework. We believe that it is because our method applies the manifold mixup technique to use the visual information more effectively. In addition, JSD plays a role in regularization, which can further improve the robustness of the model.

(2) The number of parameters in our proposed approach is slightly more than that of Transformer. More encouragingly, it is significantly lower than that of some other MMT frameworks, such as Imagination (Elliott and Kádár, 2017), which can further verify the superiority of our framework.

5 Analysis

5.1 Effect of Different Degradation Schemes

In our method, three different degradation schemes are introduced to generate these incomplete textual inputs, including x^1, x^2, and x^3, respectively. To verify the effectivenss of the degradation schemes, we sequentially remove these degradation schemes and accordingly modify the Eq. 13, denoting them as "*w/o COM*", "*w/o CHM*", and "*w/o NM*", respectively. These experimental results are demonstrated in Table 3. We could observe varying degrees of decline, indicating that all three degradation schemes play a positive role. Besides, we also attempt to replace these three degradation schemes with random masking, denoted as "*+ RM*". It can be observed that the proposed masking strategies are more effective. We believe that the reason is that via utilizing these mask strategies, the masked words could have a stronger correlation with the image, thereby further enhancing the ability to use visual information.

5.2 Ablation Study

To verify the advantages of Soul-Mix from different perspectives, we conduct ablation experiments in English-to-German translation direction. Experiment results on Test2016 dataset, Test2017 dataset, and MSCOCO dataset are shown in Table 4.

(1) **Effectiveness of Manifold Mixup**. One of the core contributions of Soul-Mix is to apply the Manifold Mixup technique to mix up the predicted translations of complete textual input and limited textual inputs. To verify the effectiveness of manifold mixup, we remove $\mathcal{L}_{\text{CE}}^2$ and $\mathcal{L}_{\text{JSD}}$ in Eq.13 and refer it to *w/o mixup* in Table 4. We could observe the performance degradation. These results demonstrate that manifold mixup can indeed improve the performance of the framework through leveraging the visual information more effectively.

(2) **Effectiveness of JSD**. To validate the effectiveness of JSD, we remove $\mathcal{L}_{\text{JSD}}$ in Eq.13 and refer it to *w/o JSD* in Table 4. We could find that BLEU drops by 0.37, 0.12, and 0.48 on these three datasets, respectively. In addition, METEOR also

SRC[En]: A brown dog about to catch a green Frisbee.
REF[De]: Ein brauner Hund ist kurz davor, einen Frisbee zu fangen.
VALHALLA(M)[De]: Ein ~~grau~~ Hund ist kurz davor, einen Platten zu fangen.
(A ~~grey~~ dog about to catch a green plate.)
E2H-MNMT[De]: Ein ~~grau~~ Hund ist kurz davor, einen Frisbee zu fangen.
(A ~~grey~~ dog about to catch a green Frisbee.)
Soul-Mix[De]: Ein brauner Hund ist kurz davor, einen Frisbee zu fangen. ✓
(A brown dog about to catch a green Frisbee.) ✓

Table 5: Case study on the Test2016 dataset of English-to-German translation direction. ~~Strikethrough~~ words present the incorrect choices. Underline denotes the acceptable but not totally right translation.

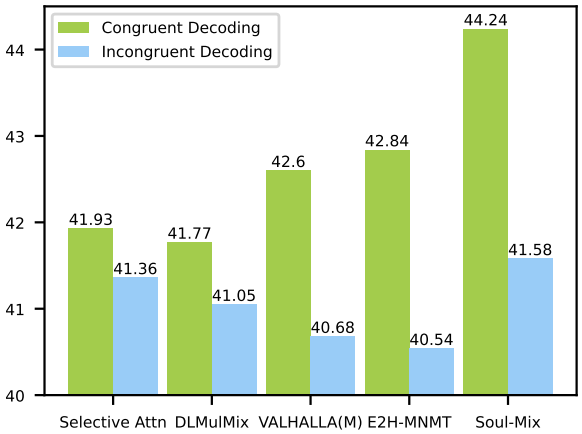

Figure 3: BLEU scores on the Test 2016 test set of the English-to-German translation direction when utilizing congruent decoding and incongruent decoding.

drops by 0.22, 0.33, and 0.49, respectively. These results show that JSD could indeed further enhance the performance of MMT. We believe that it is because JSD could act as a regularization technique to prevent overfitting during the training process.

5.3 Incongruent Decoding

Incongruent decoding involves replacing the original image with an incongruent one during the test process. The MMT framework that better leverages visual information typically experiences more performance degradation with incongruent decoding.

As shown in Figure 3, we employ incongruent decoding on the English-to-German translation direction of the Test2016 test set to test whether our method can effectively integrate the visual modality. We could clearly observe that Soul-Mix shows greater performance degradation compared to other frameworks, further demonstrating its superior ability to effectively leverage visual information.

5.4 Case Study

Finally, we perform a case study to further demonstrate the advantages of our framework. We select a text and its corresponding image from the Test2016 dataset for the English-to-German translation direction. We choose two MMT frameworks for comparison, including VALHALLA(M) (Li et al., 2022b) and E2E-MNMT (Ye et al., 2023).

As illustrated in Table 5, we could obviously observe that VALHALLA(M) fails to incorrectly predict "Frisbee" as "plate", while both E2E-MNMT and Soul-Mix correctly predict it. We believe the reason is that E2E-MNMT and Soul-Mix successfully leverage visual information to generate more accurate translations, especially for the words with multiple meanings. Besides, we can also find that VALHALLA(M) and E2E-MNMT incorrectly predict "brown" as "grey", while Soul-Mix correctly predicts it. A possible reason is that the two frameworks are misled by the grey tail of the dog in the picture, and they mistakenly regard the color of the dog as "brown". These results further validate that Soul-Mix could avoid being misled by visual information more effectively than previous frameworks.

6 Conclusion

In this paper, we propose Soul-Mix to leverage visual information more effectively via the manifold mixup technique for MMT. Experiment results on the Multi30K dataset of three translation directions demonstrate the effectiveness and the superiority of our proposed framework, which can surpass all the previous works with a relatively small increase in the number of parameters. Further analysis shows that our framework is more robust to the noise from irrelevant visual information. In the future, we will explore how to apply the manifold mixup technique to visual inputs for MMT more effectively.

Limitations

While our Soul-Mix model has achieved the encouraging performance, it is important to recognize that our current method is predicated on the utilization

of image data during inference. This indicates that the efficacy of our method is dependent on the presence of visual data. Moving forward, our research agenda will focus on overcoming this constraint by investigating new techniques to boost our model's performance when only the textual data is accessible during the inference phase.

Ethics Statement

All the data utilized in this study is gathered from the public resources. While our method has demonstrated the promising performance, it is important to note that the results they produce are not guaranteed to be flawless. Therefore, users should exercise caution and not solely depend on these results.

Acknowledgements

This paper was partially supported by NSFC (No: 62176008). We thank all the anonymous reviewers for their insightful comments.

References

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *ArXiv preprint*.

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*.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proc. of EACL*.

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

Zefan Cai, Baobao Chang, and Wenjuan Han. 2023. Human-in-the-loop through chain-of-thought. *ArXiv preprint*.

Zefan Cai, Po-Nien Kung, Ashima Suvarna, Mingyu Derek Ma, Hritik Bansal, Baobao Chang, P Jeffrey Brantingham, Wei Wang, and Nanyun Peng. 2024. Improving event definition following for zero-shot event detection. *ArXiv preprint*.

Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. 2022. Locvtp: Video-text pre-training for temporal localization. In *Proc. of ECCV*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proc. of ACL*.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2024. On the pareto front of multilingual neural machine translation. *Proc. of NeurIPS*.

Liang Chen, Runxin Xu, and Baobao Chang. 2022. Focus on the target's vocabulary: Masked label smoothing for machine translation. In *Proc. of ACL*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Proc. of ACL Findings*.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023b. M 3 st: Mix at three levels for speech translation. In *Proc. of ICASSP*.

Xuxin Cheng, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023c. C 2 a-slu: cross and contrastive attention for improving asr robustness in spoken language understanding. In *Proc. of Interspeech*.

Xuxin Cheng, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023d. Das-cl: Towards multimodal machine translation via dual-level asymmetric contrastive learning. In *Proc. of CIKM*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proc. of ACL*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Guanting Dong, Tingfeng Hui, Zhuoma GongQue, Jinxu Zhao, Daichi Guo, Gang Zhao, Keqing He, and Weiran Xu. 2023a. Demonsf: A multi-task demonstration-based generative framework for noisy slot filling task. In *Proc. of EMNLP Findings*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023b. How abilities in large language models are affected by supervised fine-tuning data composition. *ArXiv preprint*.

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*.

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*.

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)*.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proc. of ACL*.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proc. of ACL*.

Shangtong Gui, Chenze Shao, Zhengrui Ma, Yunji Chen, Yang Feng, et al. 2024. Non-autoregressive machine translation with probabilistic context-free grammar. *Proc. of NeurIPS*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proc. of ACL*.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proc. of ACL*.

Xin Huang, Jiajun Zhang, and Chengqing Zong. 2023. Contrastive adversarial training for multi-modal machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Chuhao Jin, Yutao Zhu, Lingzhen Kong, Shijie Li, Xiao Zhang, Ruihua Song, Xu Chen, Huan Chen, Yuchong Sun, Yu Chen, et al. 2023. Joint semantic and strategy matching for persuasive dialogue. In *Proc. of EMNLP Findings*.

Amit Jindal, Arijit Ghosh Chowdhury, Aniket Didolkar, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Augmenting NLP models using latent feature interpolations. In *Proc. of COLING*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. On vision features in multimodal machine translation. In *Proc. of ACL*.

Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. 2023a. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *Proc. of ICCV*.

Jicheng Li, Pengzhi Gao, Xuanfu Wu, Yang Feng, Zhongjun He, Hua Wu, and Haifeng Wang. 2021. Mixup decoding for diverse machine translation. In *Proc. of EMNLP Findings*.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023b. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proc. of ICCV*.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogério Feris, David D. Cox, and Nuno Vasconcelos. 2022b. VALHALLA: visual hallucination for machine translation. In *Proc. of CVPR*.

Ivan Medennikov, Yuri Y. Khokhlov, Aleksei Romanenko, Dmitry Popov, Natalia A. Tomashenko, Ivan Sorokin, and Alexander Zatvornitskiy. 2018. An investigation of mixup training strategies for acoustic models in ASR. In *Proc. of INTERSPEECH*.

Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu. 2021. Mixspeech: Data augmentation for low-resource automatic speech recognition. In *Proc. of ICASSP*.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proc. of COLING*.

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 *Proc. of NAACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proc. of COLING*.

Turghun Tayir and Lin Li. 2024. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024. Encoder-decoder calibration for multimodal machine translation. *IEEE Transactions on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *Proc. of ICML*.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *Proc. of ICLR*.

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 *Proc. of ACL*.

Yifei Xin, Xiulian Peng, and Yan Lu. 2023a. Masked audio modeling with clap and multi-objective learning. In *Proc. of Interspeech*.

Yifei Xin, Dongchao Yang, and Yuexian Zou. 2022. Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification. In *Proc. of Interspeech*.

Yifei Xin, Dongchao Yang, and Yuexian Zou. 2023b. Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss. In *Proc. of ICASSP*.

Yifei Xin and Yuexian Zou. 2023. Improving audio-text retrieval via hierarchical cross-modal interaction and auxiliary captions. In *Proc. of Interspeech*.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *Proc. of ICLR*.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *Proc. of AAAI*.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proc. of ACL*.

Junjie Ye and Junjun Guo. 2022. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*.

Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022. Noise-robust cross-modal interactive learning with Text2Image mask for multimodal neural machine translation. In *Proc. of COLING*.

Junjie Ye, Xiang Yan, Junjun Guo, and Zhengtao Yu. 2023. The progressive alignment-aware multimodal fusion with easy2hard strategy for multimodal neural machine translation.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proc. of ACL*.

Yongkang Yin, Xu Li, Ying Shan, and Yuexian Zou. 2023. Afl-net: Integrating audio, facial, and lip modalities with cross-attention for robust speaker diarization in the wild. *ArXiv preprint*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *Proc. of ICLR*.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proc. of ACL*.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *Proc. of ICLR*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhihong Zhu, Xuxin Cheng, Dongsheng Chen, Zhiqi Huang, Hongxiang Li, and Yuexian Zou. 2023a. Mix before align: towards zero-shot cross-lingual sentiment analysis via soft-mix and multi-view learning. In *Proc. of Interspeech*.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023b. Enhancing code-switching for cross-lingual slu: a unified view of semantic and grammatical coherence. In *Proc. of EMNLP*.

Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *Proc. of WSDM*.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024. Towards explainable joint models via information theory for multiple intent detection and slot filling. In *Proc. of AAAI*.

A Dataset

We employ four test sets to evaluate MMT models:

- The Test2016 (Elliott et al., 2016) test set with 1,000 text-image pairs in the initial Multi30K dataset. The languages in the Test2016 test set include English, German, French, and Czech.
- The Test2017 (Elliott et al., 2017) test set with 1,000 text-image pairs from WMT2017, with more difficult source sentences. Languages in the Test2017 test set include English, German, and French, respectively.
- The MSCOCO (Elliott et al., 2017) test set with 461 text-image pairs. Note that it is more challenging because there are more ambiguous verbs and instances. The languages in this dataset include English, German, and French.
- The Test2018 (Barrault et al., 2018) test set with 1,071 text-image pairs, including more entity words and low-frequency words. Languages in Test2018 include English, German, French, and Czech, respectively.