

Supplementary Materials

Anonymous Authors

1 EXPERIMENTS

1.1 Experimental settings

Implementation Details. We employ byte pair encoding (BPE) segmentation with 8k, 10k, and 6k merge operations for the Fashion-MMT, EMMT and Multi-30k datasets, respectively. The vocabulary sizes are 8880-2936 tokens for the Fashion-MMT dataset, 10407-9799 tokens for the EMMT dataset, 5644-5876 tokens for the Multi-30k (En-De) translation task, and 5644-5972 tokens for the Multi-30k (En-Fr) translation task. We utilize the pre-trained CLIP model [7] to represent textual and visual features into a shared multi-modal space, thereby obtaining the most semantic-related image features corresponding to its text for Fashion-MMT dataset. Furthermore, our model consists of 4 stacked encoders and 4 stacked decoders based on the Transformer-based seq2seq framework for all datasets. For Fashion-MMT and EMMT datasets, the word embedding dimension is 512. The multi-head number is 4, and the dimensions of self-attention hidden state and the feed-forward hidden state are 512, 1024, respectively. For En-De and En-Fr translation tasks in Multi30k dataset, the embedding dimension, self-attention dimension, and feed-forward dimensions are 128, 128 and 256, respectively. We employ Adam to optimize our model. The learning rates are set to 0.001, 0.0001, 0.001, 0.006 and 0.007 for the Fashion-MMT(clean), Fashion-MMT(large), EMMT, Multi30k (En-De) and Multi30k (En-Fr), respectively. The warmup step is 2000. The label smoothing value is set to 0.3, 0.3, 0.25 and 0.25 for Fashion-MMT, EMMT, En-De and En-Fr tasks. The value assigned to dropout is the same as the value used for label smoothing. We implement an early-stopping strategy if the BLUE score does not improve over 15 validation steps. The model is trained on a single RTX 3090 GPU using mixed-precision training(fp16).

1.2 Comparison results on three MNMT datasets

1.2.1 Comparison Results on Multi30k Dataset in General Domain. To further confirm the robustness of our proposed method, we conduct additional experiments on the Multi30k dataset. The results of English-to-German and English-to-French translation tasks are presented in Table 1 and Table 2. The findings are as follows: 1) In comparison to existing MNMT models, our method achieves SOTA scores in the BLEU, METEOR_C, and METEOR_N metrics on the test2016, test2017, and MSCOCO test sets. 2) Compared to our reproduced NMT and MNMT models, our approach demonstrates significant improvements under the same parameter and environment settings. Furthermore, we also conduct significance tests between our reproduced models and our approach. The significance test results indicate that our model achieves a statistically significant improvement over these models ($p\text{-value} < 0.01$). 3) Our model(G) and Our model(H) also achieve comparable results on these three test sets. This confirms the effectiveness of our virtual visual scene generation module. Furthermore, it is noteworthy that the MSCOCO test split, which includes sentences with ambiguous

verbs and out-of-domain samples from the COCO Captions dataset, is often challenging for MNMT models. However, our model performs exceptionally well on this test set, which suggests it can effectively employ visual information to handle ambiguity through our proposed modality exchanging mechanism.

1.3 Ablation Study

1.3.1 Effect of the Voting Times of the Inter-modality Exchanging Module. Then we investigate the effect of the voting times of the inter-modality exchanging module, as depicted in Table 3. The conclusions could be drawn as follows: 1) As the voting times increases from 1 to 5, our model exhibits an upward trend in both BLEU and BLEURT scores on Fashion-MMT(clean) and Fashion-MMT(large). 2) Conversely, as the voting times rises from 5 to 7, our model demonstrates a declining trend in BLEU and BLEURT scores on Fashion-MMT(clean) and Fashion-MMT(large). When $\tau = 5$, the model achieves the highest BLEU and BLEURT scores. Thus, we select $\tau = 5$ for analysis in this paper. Furthermore, the results in Table 3 demonstrate that, by applying the voting mechanism five times, the model more effectively captures the characteristics of cross-modal exchange under semantic constraints.

1.3.2 Effect of the Number of the Exchanged Tokens of the Inter-modality Exchanging Module. To investigate the effect of the number of the exchanged tokens, we conduct experiments on the Fashion-MMT dataset, as shown in Table 4. The results reveal the following insights: 1) The model exhibits steady growth in BLEU and BLEURT scores on both the Fashion-MMT(small) and Fashion-MMT(large) datasets when the number of exchanged tokens increases from 2 to 6; 2) A decline in BLEU and BLEURT scores is observed on these datasets as the number of exchanged tokens increases from 6 to 8. So we choose $t = 6$ in this paper. This indicates that exchanging six tokens at each time enables the model to more effectively capture domain-relevant multimodal textual and visual information.

1.4 Visual Analysis

1.4.1 The Visualization of Virtual Visual Scene-guided Progressive Domain-shadow Fusion. To further explore the effectiveness of our proposed virtual visual scene-guided progressive domain-shadow fusion, we visualized three examples of gating weights of different lengths from the Fashion-MMT test set, as shown in Figure 1. The weights on the left side of the figure represent the values under real visual conditions, while the right side depicts the gating weights under virtual visual conditions. It can be observed from the figure that, regardless of whether under real or virtual visual conditions, our model can precisely focus on domain representations. Terms with strong domain characteristics, such as "crops," "shirt," "grid pattern," and "burberry," all receive higher weights, further proving the effectiveness of our approach.

Table 1: Comparison results on the En→De translation task on the Multi30k dataset. The best scores are highlighted in bold. ↑ marks that the improvement achieved by our model over the best result of our reproduced NMT and MNMT models is statistically significant, with a p-value < 0.01. The MET_C and MET_N refer to the METEOR_C and METEOR_N evaluation metrics.

Models	Multi30k En→De								
	Test2016			Test2017			MSCOCO		
	BLEU	MET_C	MET_N	BLEU	MET_C	MET_N	BLEU	MET_C	MET_N
Existing MNMT Models									
DCCN[5]	39.7	56.8	-	31.0	49.9	-	26.7	45.7	-
RMMT[11]	41.45	-	68.0	32.94	-	61.7	30.0	-	56.3
MI[2]	39.40	-	68.60	34.58	-	62.40	30.61	-	56.70
IKD-MMT[6]	41.2	58.9	-	33.8	53.2	-	30.1	48.9	-
MDA[1]	42.00	59.43	-	34.08	52.54	-	30.38	49.60	-
VALHALLA(M)[4]	42.6	-	69.3	35.1	-	62.8	30.7	-	57.6
2/3-Triplet[15]	40.48	-	-	34.62	-	-	-	-	-
Latent Diffusion[13]	41.20	-	-	32.20	-	-	28.30	-	-
Enc-Dec Calibration[8]	42.0	60.2	-	33.4	53.7	-	30.0	49.6	-
Our Reproduced NMT and MNMT Models									
Transformer[9]	40.78	59.45	66.76	32.76	51.37	58.67	28.76	48.22	53.27
Multimodal self-att[12]	41.51	58.78	67.64	32.96	51.98	59.06	29.43	48.42	54.68
Gated Fusion [11]	41.55	58.64	67.61	32.87	51.87	59.24	29.59	48.71	54.82
Selective attention[3]	42.03	59.07	67.98	34.05	52.78	61.56	30.27	49.34	55.24
Our Ground-truth and Virtual Visual Model									
Our Model(G)	42.83	60.51↑	69.67↑	35.20	54.51↑	63.08	31.21↑	51.27↑	58.24
Our Model(H)	42.85↑	60.48	69.52	35.31↑	54.48	63.09↑	31.17	51.25	58.27↑

Table 2: Comparison results on the En→Fr translation task on the Multi30k dataset. The best scores are highlighted in bold. ↑ marks that the improvement achieved by our model over the best result of our reproduced NMT and MNMT models is statistically significant, with a p-value < 0.01. The MET_C and MET_N refer to the METEOR_C and METEOR_N evaluation metrics.

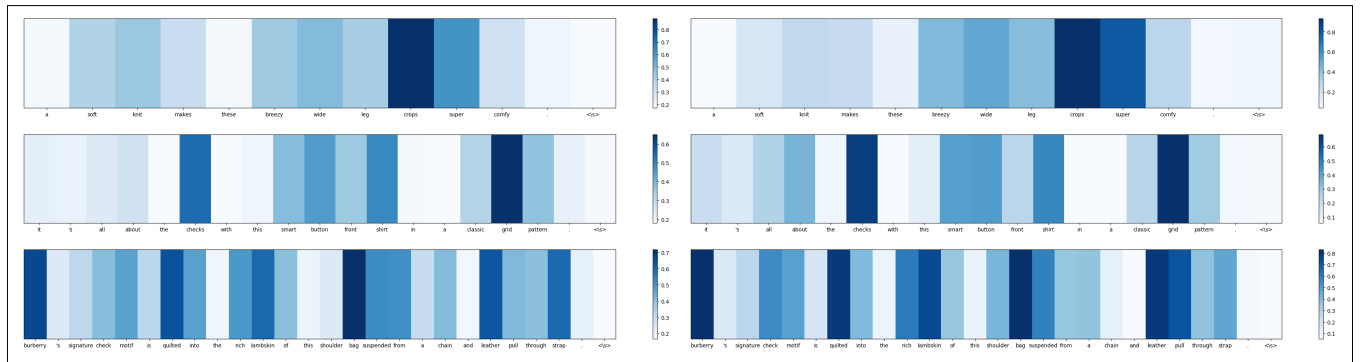
Models	Multi30k En→Fr								
	Test2016			Test2017			MSCOCO		
	BLEU	MET_C	MET_N	BLEU	MET_C	MET_N	BLEU	MET_C	MET_N
Existing MNMT Models									
DCCN[5]	61.2	76.4	-	54.3	70.3	-	45.4	65.0	-
OVC[10]	-	-	-	54.2	70.5	-	45.2	64.6	-
WRA-guided[14]	61.8	76.3	-	54.1	70.6	-	43.4	63.8	-
RMMT[11]	62.12	-	81.3	54.39	-	76.1	44.52	-	70.2
IKD-MMT[6]	62.5	77.2	-	54.8	71.8	-	-	-	-
MDA[1]	62.36	77.20	-	54.09	72.09	-	46.48	66.71	-
VALHALLA[4]	63.1	-	81.8	56.0	-	77.1	46.4	-	71.3
Enc-Dec Calibration[8]	62.9	77.2	-	55.8	72.0	-	45.1	64.9	-
Our Reproduced NMT and MNMT Models									
Transformer[9]	60.48	75.83	78.05	53.12	70.85	73.32	43.75	64.48	68.48
Multimodal self-att[12]	60.96	74.98	78.55	54.17	71.22	74.76	44.35	64.65	69.17
Gated Fusion MNMT[11]	61.46	75.27	79.25	53.93	71.34	74.94	44.21	64.26	69.06
Selective attention[3]	61.78	76.23	80.56	54.27	72.25	75.78	44.89	65.22	69.84
Our Ground-truth and Virtual Visual Model									
Our Model(G)	63.24	77.84↑	82.03↑	56.29↑	73.22↑	77.65	46.83↑	67.40	71.65
Our Model(H)	63.27↑	77.75	81.93	56.19	73.20	77.78↑	46.71	67.42↑	71.73↑

Table 3: Ablation study on the voting times in the inter-modality exchanging module.

Voting times	Model	Fashion-MMT En→Zh			
		Fashion-MMT(clean)		Fashion-MMT(large)	
		BLEU	BLEURT	BLEU	BLEURT
1	Our model(G)	41.02	59.44	43.77	61.47
	Our model(H)	41.05	59.31	43.69	61.55
2	Our model(G)	41.15	59.49	43.86	61.65
	Our model(H)	41.11	59.38	43.84	61.67
3	Our model(G)	41.33	59.67	44.14	61.89
	Our model(H)	41.29	59.71	44.11	61.73
4	Our model(G)	41.42	59.89	44.21	61.88
	Our model(H)	41.44	59.85	44.18	61.76
5	Our model(G)	41.57	60.47	44.43	62.03
	Our model(H)	41.52	60.53	44.40	62.13
6	Our model(G)	41.46	60.22	44.31	62.07
	Our model(H)	41.42	60.17	44.37	61.96
7	Our model(G)	41.33	60.01	44.06	61.52
	Our model(H)	41.34	59.89	43.98	61.36

Table 4: Effect of the number of the exchanged tokens in the inter-modality exchanging module.

The number of exchanged tokens	Model	Fashion-MMT En→Zh			
		Fashion-MMT(clean)		Fashion-MMT(large)	
		BLEU	BLEURT	BLEU	BLEURT
2	Our model(G)	41.08	59.59	43.67	61.05
	Our model(H)	41.15	59.69	43.72	61.01
3	Our model(G)	41.25	59.95	43.87	61.18
	Our model(H)	41.40	59.77	43.92	61.23
4	Our model(G)	41.35	59.89	44.05	61.22
	Our model(H)	41.31	59.93	44.08	61.41
5	Our model(G)	41.40	60.07	44.18	61.56
	Our model(H)	41.37	59.94	44.21	61.62
6	Our model(G)	41.57	60.47	44.43	62.03
	Our model(H)	41.52	60.53	44.40	62.13
7	Our model(G)	41.44	60.07	44.17	61.17
	Our model(H)	41.40	60.11	44.09	61.23
8	Our model(G)	41.21	59.78	43.76	60.01
	Our model(H)	41.17	59.75	43.79	60.06

**Figure 1: The visualization of the cross-modal gating weights for three examples of different lengths from the Fashion-MMT test set. The left and right sides denote the results of Our Model(G) and Our Model(H).**

REFERENCES

- [1] Junjun Guo, Junjie Ye, Yan Xiang, and Zhengtao Yu. 2023. Layer-level Progressive Transformer with Modality Difference Awareness for Multi-modal Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).
- [2] Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu, and Si Shen. 2022. Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective. In *2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2022.
- [3] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022. On Vision Features in Multimodal Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6327–6337.
- [4] Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5216–5226.
- [5] Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1320–1329.
- [6] Ru Peng, Yawen Zeng, and Jake Zhao. 2022. Distill The Image to Nowhere: Inversion Knowledge Distillation for Multimodal Machine Translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2379–2390.
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [8] 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* (2024).
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [10] Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 2720–2728.
- [11] 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)*. 6153–6166.
- [12] Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 4346–4350.
- [13] Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiware, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal Neural Machine Translation Using Synthetic Images Transformed by Latent Diffusion Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. 76–82.
- [14] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiware, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 244–259.
- [15] Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, and Mingxuan Wang. 2023. Beyond Triplet: Leveraging the Most Data for Multimodal Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 2679–2697.