Multimodal Neural Machine Translation Using Synthetic Images Transformed by Latent Diffusion Model

Ryoya Yuasa[1] **Akihiro Tamura**[1] **Tomoyuki Kajiwara**[2]
Takashi Ninomiya[2] **Tsuneo Kato**[1]
[1]Doshisha University [2]Ehime University
{ctwh0190@mail4, aktamura@mail, tsukato@mail}.doshisha.ac.jp
{kajiwara, ninomiya}@cs.ehime-u.ac.jp

Abstract

This study proposes a new multimodal neural machine translation (MNMT) model using synthetic images transformed by a latent diffusion model. MNMT translates a source language sentence based on its related image, but the image usually contains noisy information that are not relevant to the source language sentence. Our proposed method first generates a synthetic image corresponding to the content of the source language sentence by using a latent diffusion model and then performs translation based on the synthetic image. The experiments on the English-German translation tasks using the Multi30k dataset demonstrate the effectiveness of the proposed method.

1 Introduction

Recently, multimodal neural machine translation (MNMT) (Specia et al., 2016), which uses images in addition to source language sentences for translation, has attracted attention in the field of machine translation (MT). Images related to source language sentences are considered to improve translation performance by resolving ambiguity during translation and complementing information that is difficult to capture with source language sentences. However, a source language sentence often only describes one aspect of the contents included in its related image.

Figure 1 shows an example from a standard dataset in MNMT, the Multi30k dataset (Elliott et al., 2016). As shown in Figure 1, multiple source language sentences with differing content are associated with a single image in the Multi30k. For example, Source Language Sentence 2 does not mention the house in the related image. Therefore, related images are not necessarily optimal as auxiliary information for MT.

Therefore, in this study, we propose a new MNMT model using a synthetic image generated by image conversion with a latent diffusion model. Specifically, an original related image is converted with a latent diffusion model based on its source language sentence; content unrelated to the source language sentence is eliminated from the original image, and an image conforming with the source language sentence is generated. Subsequently, translation is performed by using the converted synthetic image instead of the original related image. Our aim is to improve translation performance by using related images that better reflect the content of source language sentences as auxiliary information for translation.

We verified the effectiveness of our proposed method on the English-German translation tasks using the Multi30k dataset (Elliott et al., 2016) and the Ambiguous COCO dataset (Elliott et al., 2017). The results confirmed that, compared with a conventional MNMT using the original related images in the Multi30k, our method improved the BLEU score by 0.14 on both the Multi30k Test 2016 and Test 2017, and by 0.39 on the Ambiguous COCO. Additionally, CLIPScore (Hessel et al., 2021), which was used to calculate the similarity between a source language sentence and an image, confirmed that the synthetic images used in our method more closely match the source language sentences than the original related images.

2 Conventional MNMT Models

MNMT models based on Transformer (Vaswani et al., 2017) have recently become mainstream in the field of MNMT. Various attempts have been made to improve their translation performance, including the introduction of visual attention mechanisms (Nishihara et al., 2020), as well as the method of simultaneously learning feature representations of text and images using a shared encoder (Elliott and Kádár, 2017). Li et al. (2022) have proposed a Transformer MNMT model incorporating Selective Attention, an attention mecha-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 76–82
July 10-12, 2023 ©2023 Association for Computational Linguistics

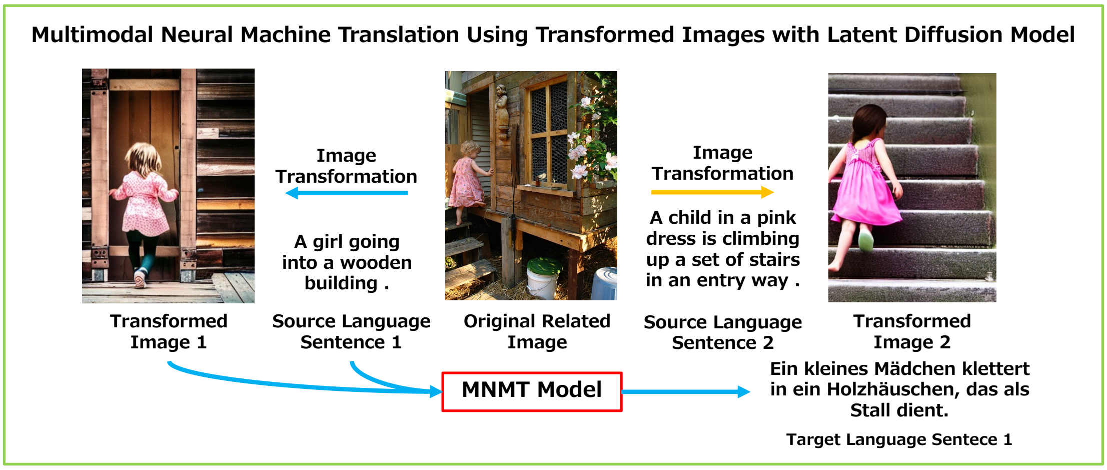
Multimodal Neural Machine Translation Using Transformed Images with Latent Diffusion Model
Image Transformation
A girl going into a wooden building .
Image Transformation
A child in a pink dress is climbing up a set of stairs in an entry way .
Transformed Image 1
Source Language Sentence 1
Original Related Image
Source Language Sentence 2
Transformed Image 2
MNMT Model
Ein kleines Mädchen klettert in ein Holzhäuschen, das als Stall dient.
Target Language Sentece 1

Figure 1: Overview of the Proposed Method

nism that captures relationships between words in a source language sentence and patches of its related image. We outline the Selective Attention MNMT model, which is used as the base MNMT model in this study, below.

The Selective Attention MNMT model first encodes the source language sentence X^{text} and the related image X^{img} into feature expressions H^{text} and H^{img} by Eqs. (1) and (2), respectively.

$$H^{\text{text}} = \text{TextEncoder}(X^{\text{text}}), \tag{1}$$

$$H^{\text{img}} = W\,\text{ImageEncoder}(X^{\text{img}}), \tag{2}$$

where W, TextEncoder, and ImageEncoder are the parameter matrix, Transformer Encoder, and Vision Transformer (Dosovitskiy et al., 2021), respectively.

Then, Selective Attention captures relationships between image patches and source words using an attention mechanism as follows:

$$H^{\text{img}}_{attn} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

where Q, K, and V are H^{text}, H^{img}, and H^{img}, respectively, and d_k is the dimension of H^{text}.

Subsequently, the gated fusion mechanism (Zhang et al., 2020) generates a feature expression H^{out} that represents the source language sentence and the image while controlling the influence of the image by Eqs. (4) and (5).

$$\lambda = \text{Sigmoid}(UH^{\text{text}} + VH^{\text{img}}_{attn}), \tag{4}$$

$$H^{\text{out}} = (1-\lambda)\cdot H^{\text{text}} + \lambda \cdot H^{\text{img}}_{attn}, \tag{5}$$

where U and V are learnable parameter matrices. Finally, H^{out} is input to the Transformer Decoder to generate a translated sentence.

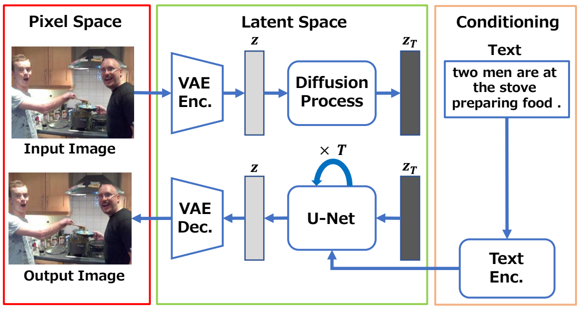

Figure 2: Training Process of a Latent Diffusion Model

3 Proposed Method

In this section, we propose an MNMT model that uses synthetic images transformed from related images based on source language sentences. Figure 1 shows an overview of the proposed method.

The MNMT dataset consists of the triplets of a source language sentence, a target language sentence, and a related image. In typical MNMT datasets, each source language sentence usually only represents one aspect of the content included in the related images; there are many cases where content unrelated to the source language sentence exists in the related image. For example, the image in Figure 1 shows a scene where a girl in a pink dress climbs the stairs to enter a wooden house, but Source Language Sentence 1 does not mention the climbing of stairs. Further, Source Language Sentence 2 does not refer to a house. Therefore, related images are not necessarily the best aids to translation.

Accordingly, our proposed method first uses a latent diffusion model to eliminate content unrelated to the source language sentence from the related image and generate a synthetic image that corresponds to the source language sentence (see Section 3.1). Then, translation is performed with a conventional MNMT model (e.g., the Selective Attention MNMT model in our experiments) using the generated synthetic image and the source language sentence. Because this makes it easier to capture the relationship between the input image and text during translation, we expect the improvement of translation performance.

3.1 Image Transformation: Latent Diffusion Model

This section explains the latent diffusion model (Rombach et al., 2022) used in the image transformation step of our proposed method. The latent diffusion model applies the diffusion model (Sohl-Dickstein et al., 2015) to the latent space of VAE (Kingma and Welling, 2014) and consists mainly of the VAE, U-Net (Ronneberger et al., 2015), and a text encoder (see Figure 2). In the latent diffusion model, an input image is projected from pixel space into a low-dimensional latent space using a VAE Encoder to obtain its latent representation. Then Gaussian noise is continuously added to the latent expression by a diffusion process. Next, in a reverse diffusion process, U-Net is used multiple times to gradually remove noise from the latent expression that contained noise. At this time, the U-Net is conditioned by the feature representation generated from a text by the text encoder. This conditioning is realized by a cross attention mechanism. Finally, the VAE decoder projects the denoised latent representation from latent space to pixel space to obtain the output image.

The loss function for the latent diffusion model is given as follows:

$$L_{\text{LDM}} := \mathbb{E}_{\varepsilon(x),y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2],$$

where ε, ϵ_θ, and τ_θ represent a VAE encoder, an U-Net, and a text encoder, respectively, and x, y, ϵ, t, and z_t are an input image, a text, a Gaussian noise, time, and the latent representation of time t, respectively.

In our proposed method, a source language sentence and its related image are input to the text encoder and the VAE encoder, respectively, to convert the related image into a synthetic image that conforms to the source language sentence.

4 Experiments

4.1 Experimental Setup

We verified the effectiveness of the proposed method on the English-German translation tasks using the Multi30k and the Ambiguous COCO. We used the Multi30k training data (29,000 triplets) and the Multi30k validation data (1,014 triplets) as our training and validation data, and used the Multi30k Test 2016 (1,000 triplets), the Multi30k Test 2017 (1,000 triplets), and the Ambiguous COCO (461 triplets) as our test data.

We compared the translation performance of our proposed method (*MNMT(conv.)*) with the translation performance of 1) an NMT model that does not use related images (*NMT*); 2) an MNMT model that uses original images from the dataset as related images (*MNMT(orig.)*); 3) and an MNMT model that uses images generated only from source language sentences as related images (*MNMT(gen.)*).

Transformer-Tiny[1] was used as the NMT model. This model, with a reduced number of layers, size of hidden layers, number of attention mechanism heads, etc., as compared to typical Transformer models, is suitable for small-scale datasets.[2] According to Wu et al. (2021), we set the number of encoder and decoder layers, the size of the hidden layer, the input size of the feed-forward layer, the number of attention mechanism heads, the dropout, and the label smoothing weight to 4, 128, 256, 5, 0.3, and 0.1, respectively. Adam (Kingma and Ba, 2015) was used as the optimization method, with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate was linearly warmed up from $1e^{-7}$ to $5e^{-3}$ over the first 2,000 steps, and then it was decreased proportionally to the number of updates. The vocabulary dictionary was shared between the source language and the target language, and created by Byte Pair Encoding (Sennrich et al., 2016) with 10,000 merge operations.

The Selective Attention MNMT[3] was used as the MNMT model. As for Vision Transformer, vit_base_patch16_384[4] was used for image feature extraction. Stable Diffusion,[5] based on a latent

[1] https://github.com/LividWo/Revisit-MMT

[2] Wu et al. (2021) reported that Transformer-Tiny outperforms Transformer Base/Small on the Multi30k dataset.

[3] https://github.com/libeineu/fairseq_mmt

[4] https://github.com/rwightman/pytorch-image-models

[5] https://github.com/CompVis/

Model	Test	Test	Ambiguous
	2016	2017	COCO
NMT	40.50	31.31	27.81
MNMT(orig.)	41.06	32.06	27.91
MNMT(gen.)	40.81	31.81	**28.54**
MNMT(conv.)	**41.20**	**32.20**	28.30

Table 1: Translation Performance (BLEU [%])

Model	Test	Test	Ambiguous
	2016	2017	COCO
MNMT(orig.)	79.59	78.32	78.17
MNMT(conv.)	79.74	79.35	80.08

Table 2: CLIPScore: Similarity between Source Language Sentences and Related Images

diffusion model, was adopted for the generation of related images in *MNMT(gen.)* and the image transformation in *MNMT(conv.)*; the specific model used was stable-diffusion-v1-5.[6] StableDiffusionPipeline and StableDiffusionImg2ImgPipeline from diffusers,[7] were used for implementation. For image generation in *MNMT(conv.)* and *MNMT(gen.)*, we used the default parameters. We set guidance_scale and num_inference_steps to 7.5 and 50 for *MNMT(gen.)*, and guidance_scale and strength to 7.5 and 0.8 for *MNMT(conv.)*. The hyperparameters, optimization methods, and vocabulary dictionary creation methods during training were the same as the settings used for the NMT model.

In decoding for all models, we averaged checkpoints at the last 10 epochs before the end of training, and used beam search with a beam width of 5. BLEU (Papineni et al., 2002) was used as the evaluation measure. We trained the models with five different random seeds, and evaluated the model with the highest BLEU on the validation data.

4.2 Results

Table 1 shows the experimental results. As Table 1 shows, the three MNMT models using image information have higher BLEU scores across all datasets than the NMT model that does not use image information. This confirms that image information helped improve translation performance on the datasets used in our experiments.

Further, a comparison of the three MNMT models shows that our proposed *MNMT(conv.)* achieved the highest translation performance on Test 2016 and Test 2017. *MNMT(gen.)* had a higher translation performance than *MNMT (conv.)* on Ambiguous COCO, but overall, *MNMT (conv.)* had better results, confirming the effectiveness of the proposed method.

5 Discussion

This section analyzes the synthetic images used in the proposed method. Examples of transformed images are shown in Appendix A. In order to investigate how much of the image corresponds to the source language sentence, we computed ClipScore (Hessel et al., 2021), which measures the similarity between the image used and the source language sentence by using CLIPScore$(\boldsymbol{c}, \boldsymbol{v}) = w \cdot \max(\cos(\boldsymbol{c}, \boldsymbol{v}), 0)$, where $\boldsymbol{c}$ and $\boldsymbol{v}$ are the feature vectors from the text encoder and the image encoder of the CLIP (Radford et al., 2021), respectively. w is used to rescale the output, and following Hessel et al. (2021), we set it to 2.5.

The evaluation results are shown in Table 2. The table shows that the synthetic images converted by our proposed method have a higher similarity to the source language sentences than the original related images across all datasets. In particular, the largest improvement (+1.91 CLIPScore) has been observed on Ambiguous COCO, which includes more ambiguity than the other two test datasets. These results confirm that related images which better reflect the source languages can be used as aids to translation via our proposed method.

6 Conclusion

In this study, we proposed a new MNMT model that uses a latent diffusion model to transform related images into synthetic images that more closely conform to source language sentences and uses the transformed images as auxiliary information for MT. The experiments on the English-German translation tasks using the Multi30k dataset showed that the proposed method can achieve higher translation performance than conventional methods, demonstrating the effectiveness of our proposed method. The evaluation using CLIPScore confirms that the images used in our method possess more similarities to the source language sentences than the original images.

stable-diffusion

[6] https://huggingface.co/runwayml/stable-diffusion-v1-5

[7] https://github.com/huggingface/diffusers

Limitations

In this work, we confirm the effectiveness of the proposed method only on the English-German translation tasks using the Multi30k dataset, the most commonly used dataset in the MNMT reserach area. It is not clear whether the proposed method is effective for translation for language pairs other than English and German or translation when a larger training dataset is used (e.g., when using an existing data augmentation method for MNMT). We will leave these verification experiments for future work.

The proposed method has improved translation performance of MT, but the performance is not perfect and translation results could include translation errors. Accordingly, there still remains a possibility that translation results by the proposed method could convey incorrect information.

The proposed method requires an additional process for transforming images, compared with conventional MNMT models. The experiment, including model training and testing, on the proposed model *MNMT(conv.)* took about 20 hours longer than that on the baseline MNMT model *MNMT(orig.)* when using RTX3090 GPU × 1.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP22K12177, JP21K12031. These research results were partially obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), JAPAN.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (Poster)*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*.

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)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314, Barcelona, Spain (Online). International Committee on 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. 2021. 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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference* *on Computer Vision and Pattern Recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265.

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.

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.

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

A Appendix

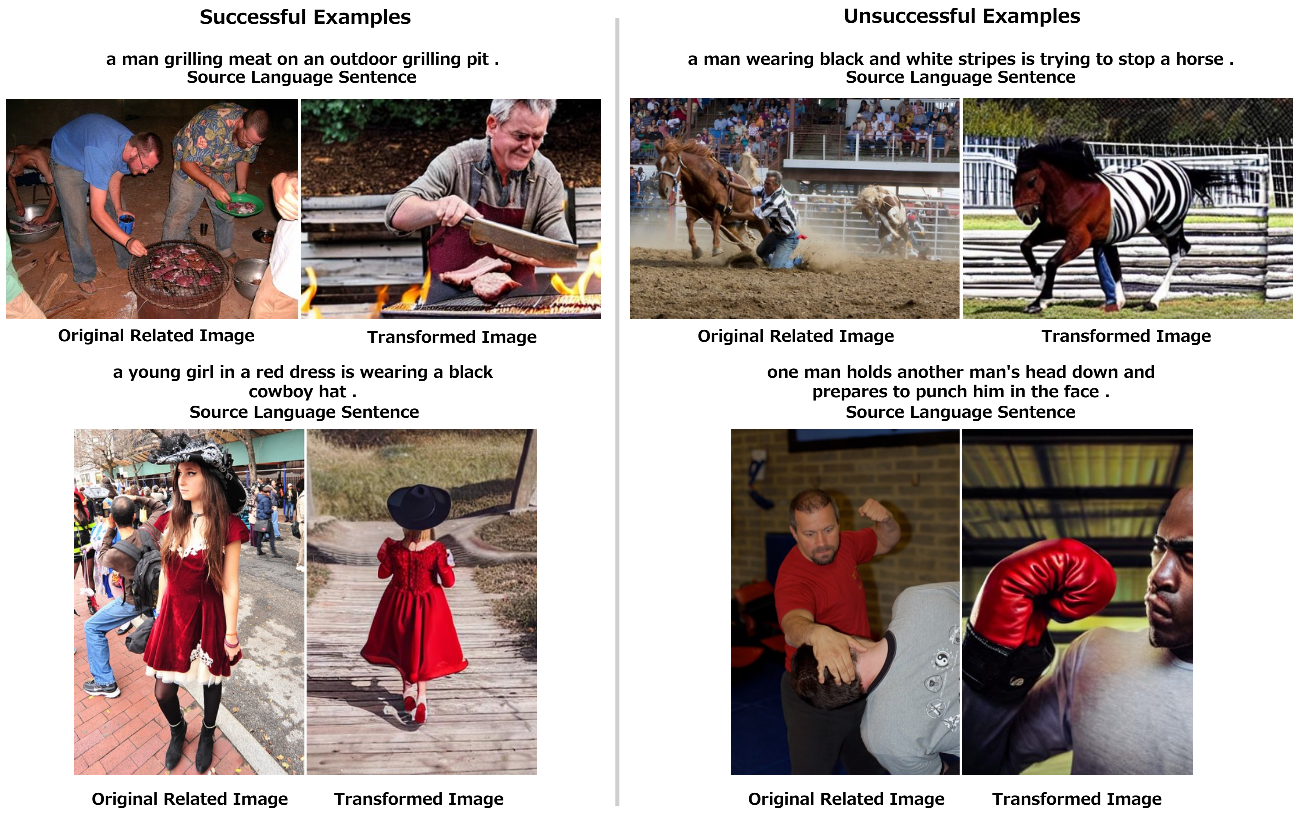

Figure 3: Successful (Left) and Unsuccessful (Right) Examples of our Image Transformation