A IKEA Dataset Stats and Examples

We summarize the statistics of our IKEA dataset in Figure 1, where we demonstrate the information about the number of tokens, the length of the product description and the vocabulary size. We provide one example of the IKEA dataset in Figure 2.

Pair	EN-DE		EN-FR	
Language	EN	DE	EN	FR
Tokens	256355	216892	239966	275251
Min length	6	6	6	6
Max length	343	324	334	469
Avg length	71.4	60.4	72.2	82.9
Std dev	46.3	39.1	47.2	54.7
Vocabulary	6601	10468	6442	7575

Figure 1: Statistic of the IKEA dataset

B Hyperparameter Settings

In this Appendix, we share details on the hyperparameter settings for our model and the training process. The word embedding size for both encoder and decoder are 256. The Encoder is a one-layer bidirectional recurrent neural network with Gated Recurrent Unit (GRU), which has a hidden size of 512. The decoder is a recurrent neural network with conditional GRU of the same hidden size. The visual representation is a 2048-dim vector extracted from the `pool5` layer of a pre-trained ResNet-50 network. The dimension of the shared visual-text semantic embedding space is 512. We set the decoder initialization weight value λ to 0.5.

During training, we use Adam (Kingma and Ba, 2014) to optimize our model with a learning rate of $4e-4$ for German Dataset and $1e-3$ for French dataset. The batch size is 32. The total gradient norm is clipped to 1 (Pascanu et al., 2012). Dropout is applied at the embedding layer in the encoder, context vectors extracted from the encoder and the output layer of the decoder. For Multi30K German dataset the dropout probabilities are $(0.3, 0.5, 0.5)$ and for Multi30K French dataset the dropout probabilities are $(0.2, 0.4, 0.4)$. For the Multimodal shared space learning objective function, the margin size γ is set to 0.1. The objective split weight α is set to 0.99. We initialize the weights of all the parameters with the method introduced in (He et al., 2015).

(a) *Product image*

the 4 large drawers on casters give you an extra storage space under the bed . adjustable bed sides allow you to use mattresses of different thicknesses . 17 slats of layer glued birch adjust to your body weight and increase the suppleness of the mattress . wipe clean using a damp cloth and a mild cleaner . wipe dry with a clean cloth.

(b) *Source description*

die 4 geräumigen schubladen auf rollen sorgen für zusätzlichen stauraum unter dem bett . durch verstellbare bettseiten können matratzen in verschiedenen stärken verwendet werden . mit feuchtem tuch (evtl . mit mildem reinigungsmittel) abwischen . mit trockenem tuch nachwischen .

(c) *Target description in German*

Figure 2: An example of product description and the corresponding translation in German from the **IKEA dataset**. Both descriptions provide an accurate caption for the commercial characteristics of the product in the image, but the details in the descriptions are different.

C Ablation Analysis on Visual-Text Attention

We conducted an ablation test to further evaluate the effectiveness of our visual-text attention mechanism. We created two comparison experiments where we reduced the impact of visual-text attention with different design options. In the first experiment, we remove the visual-attention mechanism in our pipeline and simply use the mean of the encoder hidden states to learn the shared embedding space. In the second experiment, we initialize the decoder with just the mean of encoder hidden states without the weighted sum of encoder states using the learned visual-text atten-

tion scores.

We run both experiments on Multi30K German Dataset five times and demonstrate the results in table 1. As can be seen, the performance of the changed translation model is obviously worse than the full VAG-NMT in both experiments. This observation suggests that the visual-attention mechanism is critical in improving the translation performance. The model improvement comes from the attention mechanism influencing the model's objective function and decoder's initialization.

	English → German	
Method	BLEU	METEOR
-attention in shared embedding	30.5 ± 0.6	51.7 ± 0.4
-attention in initialization	30.8 ± 0.8	51.9 ± 0.5
VAG-NMT	$\mathbf{31.6 \pm 0.6}$	$\mathbf{52.2 \pm 0.3}$

Table 1: Ablation analysis on visual-text attention mechanism in the Multi30K German dataset.

References

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.