VALHALLA: Visual Hallucination for Machine Translation Supplemental Material

Yi Li[1] Rameswar Panda[2] Yoon Kim[3] Chun-Fu (Richard) Chen[2]
Rogerio Feris[2] David Cox[2] Nuno Vasconcelos[1]

[1]UC San Diego
[2]MIT-IBM Watson AI Lab
[3]MIT CSAIL

http://www.svcl.ucsd.edu/projects/valhalla

Section	Content
A	Dataset Details
B	Implementation Details
C	Additional Results
D	Limitations
E	Broader Impact

Table 1: **Supplementary Material Overview.**

A. Dataset Details

We evaluate the performance of our proposed approach (**VALHALLA**) using three machine translation datasets, namely Multi30K [3], Wikipedia Image Text (WIT) [14] and WMT2014 [1]. These datasets present a diversity of challenges in machine translation: Multi30K requires models to learn to aggregate vision-language information from a relatively small number of training samples, while WIT and WMT contains translation tasks with different data scales. WMT additionally focuses on translating news articles, which may not be as readily grounded through visual data (compared to Multi30K and WIT), and thus presents an especially challenging test bed for MMT systems. Below we provide more details on each of the dataset.

A.1. Data Preprocessing

Table 2 summarizes the list of all machine translation tasks. We use byte-pair encoding (BPE) [5, 13] to tokenize all source and target sentences[1], with vocabulary size provided in the last row of the table. All sentences are preprocessed and cleaned using standard scripts[2].

[1]https://github.com/rsennrich/subword-nmt
[2]https://github.com/moses-smt/mosesdecoder

Multi30K. This is a multilingual translation dataset with 29000 training samples of images and their annotations in English, German, French and Czech. Each English description is manually translated to German by a professional translator, then expanded to French and Czech. We use English-German (EN→DE) and English-French (EN→FR) for our experiments. Besides showing results on Test2016 and Test2017 sets, we use MSCOCO for evaluation which is a small dataset collected in WMT2017 multimodal machine translation challenge for testing out-of-domain performance of translation models. This evaluation set includes 461 more challenging out-of-domain instances with ambiguous verbs.

WIT. We construct multimodal translation datasets for 7 language pairs from WIT [14] data. Sentence-image data for MMT are obtained from *reference descriptions* of the dataset, i.e., captions which are visible on the wiki page directly below the image. We empirically find these to contain richest visually grounded concepts compared to other types of captions provided in WIT. First, we generate raw ground-truth translation pairs by sampling from images with captions annotated in both source and target languages. For images associated with multiple captions in the same language, we sample one sentence at random. Finally, a cleaning process filters out noise by ranking the sentence pairs by their length ratios S/T, and discarding top and bottom 5% samples.

The validation and test splits for the original WIT are not publicly available, so we partition the training data to construct new splits for WIT with sizes provided in table 2.

WMT. We use the official train, validation and test data for standard WMT tasks. Under-resourced variants are created by downsampling the training sets of EN→DE and EN→FR tasks by approximately 3×10^{-2} and 3×10^{-3} respectively, creating subsets of 100k samples each. Validation and test

Dataset	Multi30K [3]		WIT [14]							WMT2014 [1]			
Visual Data	Flickr30K [17]									Flickr30K [17] / WIT [14]			
Source Language	EN	EN	EN	EN	EN	DE	ES	EN	EN	EN	EN	EN	EN
Target Language	DE	FR	DE	ES	FR	ES	FR	RO	AF	DE	FR	DE	FR
# Train Samples	29k	29k	329k	287k	234k	133k	122k	40k	18k	3.9m	36m	100k	100k
# Validation Samples	1k	1k	15k	15k	15k	10k	10k	5k	5k	39k	27k	39k	27k
# Test Samples	2.5k	2.5k	3k	3k	3k	2k	2k	1k	1k	3k	3k	3k	3k
BPE Vocabulary Size	10k							2k		40k		10k	

Table 2: **Datasets and Tasks**. We use 3 datasets with total 13 tasks that covers various languages with different scales of training data.

Dataset	Multi30K			WIT			WMT	
Task	All			Well-Res.	Non-English	Under-Res.	Well-Res.	Under-Res.
Model	Base	Small	Tiny	Base	Small		Base	Small
Architecture								
Enc./Dec. Layers	6	4	4	6	4		6	3
Embedding Dim.	512	256	128	512	256		512	512
Feedforward Dim.	2048	256	256	2048	256		2048	1024
Attn. Heads	8	8	4	8	8		8	8
Optimization								
Iters. / Warm-up	20k / 2k			50k / 8k			150k / 8k	40k / 4k
Batch Size (Tokens)	2048			4096			16384	8192
Learning Rate	0.0001	0.0005	0.0025	0.0005			0.0005	0.001
Dropout	0.5	0.5	0.3	0.3	0.3	0.5	0.1	0.3

Table 3: **Model Architectures and Optimization Hyperparameters.** Hyperparameters are selected by grid search on the respective validation set. Note that our *Small* model in Multi30K is different from that used by Wu et al. [16].

sets kept the same as full WMT.

Images. Discrete visual encoders (VQGAN VAE) are trained on images randomly cropped and resized to 128 pixels, with pixel values rescaled to $[0, 1]$. At test time, we use center cropping for all images instead.

A.2. Licenses

All the datasets considered in this work are publicly available. WIT[3] [14] is available under the CC BY-SA 3.0 license. Licenses for WMT 2014[4] [1] and Multi30K[5] [3] are unknown. Use of images from Flickr30K[6] [17] are subject to Flickr terms of use[7].

B. Implementation Details

In this section, we provide more implementation details regarding model architectures, training, inference procedures and hyperparameter selections.

B.1. Model Architecture

In Table 3 we provide the detailed architectures of all translation models $\mathbf{f_T}$ used for each dataset. For all experiments, we use a hallucination transformer $\mathbf{f_H}$ of depth 2 and VQGAN VAE visual encoder $\mathbf{f_V}$ of encoder depth 6.

Sinusoidal positional embeddings (PE) [15] are added to the multimodal input sequence to the translation transformer $\mathbf{f_T}$. Using a learnable PE [6] did not improve translation performance in preliminary experiments. For visual tokens, we follow [12] to compute 2D positional encoding as the sum of row and column embeddings.

B.2. Training Procedure

For all models and tasks, we optimize the **`VALHALLA`** system using Adam [7] with inverse square root learning rate schedule and warm-up steps. Table 3 lists important optimization hyperparameters used for each task and model, determined in preliminary experiments by grid search on the respective validation set.

B.3. Inference and Evaluation

During inference we use beam search with a beam size of 5 to generate translation outputs for each task. Length penalty α is set to 0.6 on full WMT dataset, 2 on WIT dataset, and 1 on all other trasnlation tasks. We use standard

[3] https://github.com/google-research-datasets/wit
[4] https://www.statmt.org/wmt14/translation-task.html
[5] https://github.com/multi30k/dataset
[6] http://hockenmaier.cs.illinois.edu/DenotationGraph/
[7] https://www.flickr.com/help/terms/

Method	Model	Params	EN → DE				EN → FR			
			Test2016	Test2017	MSCOCO	Average	Test2016	Test2017	MSCOCO	Average
Transformer-Base	T	49.1M	61.8 ± 1.3	53.3 ± 1.1	49.1 ± 1.2	54.7 ± 1.2	80.1 ± 0.3	74.5 ± 0.3	68.5 ± 0.2	74.4 ± 0.3
Transformer-Small	T	9.2M	65.6 ± 0.3	58.1 ± 0.6	52.5 ± 0.7	58.7 ± 0.5	79.2 ± 0.2	73.7 ± 0.1	67.9 ± 0.2	73.6 ± 0.2
	V	24.3M	**66.7 ± 0.4**	**60.1 ± 0.0**	**54.2 ± 0.4**	**60.3 ± 0.3**	**80.3 ± 0.2**	**74.8 ± 0.6**	68.8 ± 0.4	74.6 ± 0.4
	VM	24.3M	**66.7 ± 0.3**	**60.1 ± 0.0**	**54.2 ± 0.3**	**60.3 ± 0.2**	**80.3 ± 0.2**	74.7 ± 0.5	**69.0 ± 0.3**	**74.7 ± 0.3**
Transformer-Tiny	T	2.6M	67.8 ± 0.3	61.6 ± 0.5	56.2 ± 0.6	61.9 ± 0.5	80.6 ± 0.2	75.6 ± 0.2	69.8 ± 0.2	75.3 ± 0.2
	V	22.1M	**68.8 ± 0.2**	**62.5 ± 0.2**	57.0 ± 0.6	**62.8 ± 0.3**	**81.4 ± 0.2**	**76.4 ± 0.2**	70.9 ± 0.3	76.2 ± 0.2
	VM	22.1M	68.7 ± 0.2	**62.5 ± 0.2**	**57.2 ± 0.7**	**62.8 ± 0.4**	**81.4 ± 0.2**	**76.4 ± 0.1**	**71.0 ± 0.3**	**76.3 ± 0.2**

Table 4: **METEOR score on Multi30K**. T: Baseline text-only transformer; V: **VALHALLA** model with hallucinated visual representations; VM: **VALHALLA** model with ground-truth visual representations. Similar to BLEU score, **VALHALLA** (V) consistently outperforms the text-only baseline while being very competitive with **VALHALLA** (VM) on both tasks.

Method	Well-Resourced			Non-English		Under-Resourced		Average
	EN → DE	EN → ES	EN → FR	DE → ES	ES → FR	EN → RO	EN → AF	
Text-Only	35.4 ± 0.5	44.6 ± 1.7	37.4 ± 1.3	33.3 ± 0.3	37.0 ± 0.2	26.6 ± 0.6	30.2 ± 1.0	34.9 ± 0.8
UVR-NMT [18]	35.9 ± 0.1	46.7 ± 0.2	39.5 ± 0.5	32.7 ± 1.1	37.2 ± 0.7	28.0 ± 0.7	32.8 ± 1.4	36.1 ± 0.7
RMMT [16]	35.4 ± 0.6	44.8 ± 0.8	39.0 ± 1.0	33.2 ± 0.4	36.5 ± 0.9	23.6 ± 0.2	29.6 ± 1.3	34.6 ± 0.7
VALHALLA	**36.8 ± 0.5**	**47.1 ± 0.2**	**40.2 ± 0.3**	**34.3 ± 0.3**	**37.5 ± 0.9**	30.4 ± 0.9	**34.2 ± 0.2**	**37.2 ± 0.5**
VALHALLA (M)	36.7 ± 0.5	**47.1 ± 0.3**	**40.2 ± 0.3**	**34.3 ± 0.4**	**37.5 ± 0.9**	**30.5 ± 0.9**	**34.2 ± 0.2**	**37.2 ± 0.5**

Table 5: **METEOR score on WIT**. Our proposed, **VALHALLA** achieves an average 4 point improvement over text-only baseline in under-resource setting including best average performance among all compared methods.

scripts to compute BLEU[8] [9] and METEOR[9] [2] scores as evaluation metrics for machine translation.

B.4. Code and Models

Our **VALHALLA** framework is implemented on top of fairseq [8] using PyTorch [10]. Code and pretrained models are available at our project page: http://www.svcl.ucsd.edu/projects/valhalla.

C. Additional Results

C.1. Numerical Scores

METEOR Scores. Tables 4 and 5 summarizes METEOR scores of models evaluated on Multi30K [3] and WIT [14], respectively. Similar to the trend in BLEU scores (Tables 1 and 3 in the main paper), **VALHALLA** outperforms text-only and multimodal baselines consistently on all tasks.

Sentence Length. We repeat the study of translation performance vs. length of source sentence (Figure 3 in main paper) on Test2017 and MSCOCO evaluation sets. Figure 1 shows the results. Similar to the observations in Test2016 set, **VALHALLA** generally produces larger gains over text-only baselines for longer sentences (> 10 source tokens).

Progressive Masking. Figure 2 compares the performance of **VALHALLA** and text-only model under progressive masking, evaluated on Test2017 and MSCOCO splits. Similar to the observations in main paper (Figure 4), we observe a larger gap between **VALHALLA** and text-only model with low context sizes k, validating its effectiveness in translating ambiguous or out-of-context sentences.

Number of Parameters vs Performance. Larger model size does not guarantee stronger translation performance due to overfitting. As shown in Table 1 of main paper, among all three backbone architectures experimented on Multi30K, Transformer-Tiny with the least number of parameters achieved the highest scores, consistent with the findings of [16]. Our proposed, **VALHALLA** achieves 10 BLEU points (8 METEOR points) higher than the text-only baseline transformer on Multi30K EN→DE, and 3 BLEU points (2 METEOR points) higher on EN→FR tasks, while using $2\times$ fewer parameters.

C.2. Ablation Studies

Effect of External Data on WIT Tasks. Table 6 shows full results on WIT with a universal visual encoder pretrained on the union of images from all tasks. While this improves performance on 4 out of 7 tasks, average score over all tasks is only marginally better than individually trained encoders.

CLIP Hallucination. A naive method to predict visual features from an input sentence is to utilize a CLIP model [11], learned with a cross-modal contrastive loss that aligns the embedding space of text and image. We train a multimodal translation model with a gating strategy on top of image features extracted from the pretrained CLIP visual encoder, and replace this with text embeddings from CLIP language en-

[8] https://github.com/moses-smt/mosesdecoder/blob/RELEASE-4.0/scripts/generic/multi-bleu.perl
[9] https://github.com/cmu-mtlab/meteor

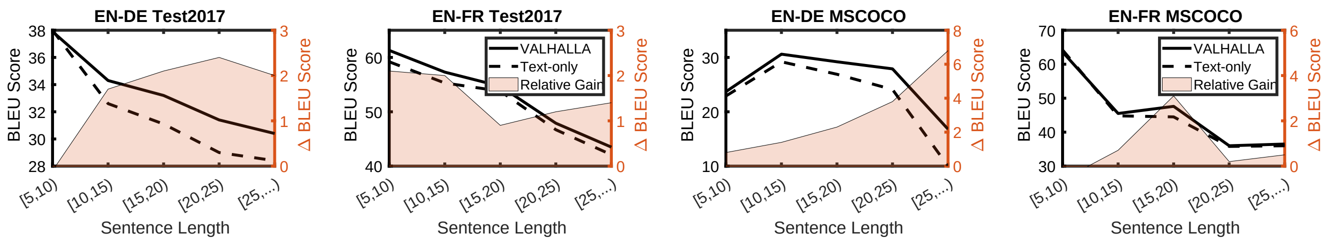


Figure 1: **Performance vs. Sentence Length.** We report BLEU scores on different groups divided according to source sentence lengths on Multi30K Test2017 and MSCOCO split.

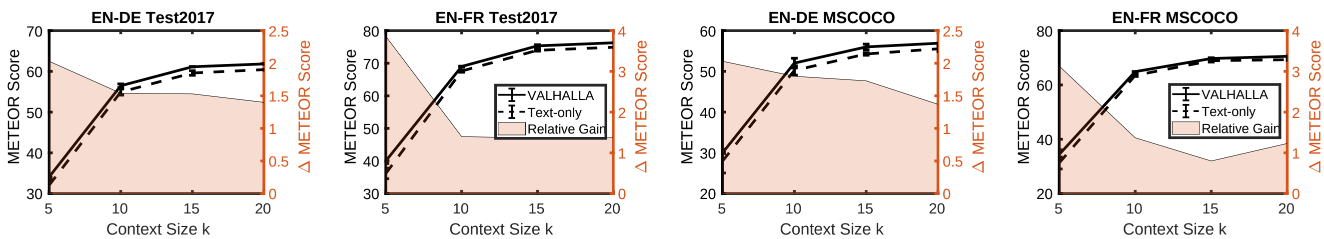


Figure 2: **Evaluation with Progressive Masking.** We plot the METEOR scores of **VALHALLA** and text-only models evaluated on Test2017 and MSCOCO splits, as well as the relative improvements over the text-only baseline on both EN→DE and EN→FR tasks.

coder to realize text-only inference. Table 7 shows the performance of this CLIP-based hallucination model on Multi30K. While CLIP-based feature hallucination consistently outperforms the text-only baseline, the improvements in BLEU score is not nearly as large as those achieved by **VALHALLA**, which still reports the best results among all strategies.

Hallucinating Multiple Images. We study the possibility of modifying the hallucination transformer $\mathbf{f_H}$ to predict multiple images for each input sentence. While this in theory enhances the diversity of hallucination, we did not observe significant improvement over baselines. By hallucinating 5 images per example, Transformer-Small model achieved 39.4 ± 0.3 and 60.4 ± 0.2 BLEU score on EN→DE and EN→FR tasks respectively, evaluated on Test2016 split. On Test2017 split, the scores are 31.8 ± 0.4 and 52.2 ± 0.1. Both results are comparable to the results reported in main paper, suggesting that a single hallucination per sample is adequate to capture diverse visual concepts in the input sentence.

Pretrained VAEs. Using the pretrained VAE from DALL-E as the visual encoder gives poor results (58.8 BLEU on Multi30K'16 EN→FR). We attribute this to the large visual sequence length (32×32) used by DALL-E, which prevents the MMT transformer to attend to text tokens, as analyzed in Table 5b of main paper. Likewise, use of pre-trained VQGAN VAE [4] with 16×16 latent visual sequence also does not improve results from training on Multi30K alone (59.5 vs. 60.5 BLEU), likely due to the larger sequence length or domain gap between pretraining datasets.

C.3. Qualitative Examples

Translation under Limited Visual Context. Figure 3 shows additional qualitative translation results under both progressive masking and visual entity masking. We observe that in both EN→DE and EN→FR tasks , our proposed **VALHALLA** models are often capable of generating more fluent and logical translations than the text-only baseline transformer, by choosing plausible phrases to replace the masked tokens in the source sentences.

Reconstructed Visual Hallucinations. Figure 4 visualizes the hallucinated visual tokens using the VQGAN VAE decoder, which is pretrained jointly with VAE encoder $\mathbf{f_V}$. As seen from the examples, **VALHALLA** captures abstract concepts such as "surfer" and "red ribbons", despite not being trained for high-quality image generation.

D. Limitations

Effectiveness of our approach depends on availability of good quality images to train the visual hallucination transformer, which is often difficult to collect especially for languages beyond English. Another potential limitation is training complexity which we believe could be greatly improved if we pre-extract VQGAN-VAE tokens, like existing methods did with ResNet-based visual encoders.

E. Broader Impact

Our approach not only leads to more accurate translation systems on top of the existing text-only methods, but also breaks the major bottleneck of using visual information in

Method	External Data	EN→DE	EN→ES	EN→FR	DE→ES	ES→FR	EN→RO	EN→AF	Average
VALHALLA	✗	17.5 ± 0.4	27.5 ± 0.2	18.8 ± 0.2	**11.3 ± 0.2**	16.6 ± 0.8	**14.4 ± 1.0**	**14.0 ± 0.5**	17.2 ± 0.4
	✓	**18.0 ± 0.3**	**27.7 ± 0.4**	**19.1 ± 0.3**	**11.3 ± 0.7**	**17.4 ± 0.4**	14.1 ± 0.3	13.8 ± 0.3	**17.3 ± 0.4**
VALHALLA (M)	✗	17.4 ± 0.4	27.5 ± 0.2	18.8 ± 0.2	**11.3 ± 0.2**	16.6 ± 0.8	**14.4 ± 1.0**	**14.0 ± 0.4**	17.2 ± 0.4
	✓	**18.0 ± 0.3**	**27.8 ± 0.4**	**19.1 ± 0.3**	**11.3 ± 0.7**	**17.4 ± 0.4**	13.9 ± 0.4	13.9 ± 0.4	**17.3 ± 0.4**

Table 6: **Training Visual Encoder (VQGAN VAE) with External Data on WIT.**

Method	EN → DE				EN → FR			
	Test2016	Test2017	MSCOCO	Average	Test2016	Test2017	MSCOCO	Average
Text-Only	38.2 ± 0.4	28.8 ± 0.4	25.8 ± 0.3	30.9 ± 0.4	58.4 ± 0.4	50.9 ± 0.3	41.6 ± 0.4	50.3 ± 0.4
CLIP	38.7 ± 0.2	30.1 ± 0.3	27.3 ± 0.6	32.1 ± 0.3	59.0 ± 0.6	51.6 ± 0.3	42.6 ± 0.6	51.1 ± 0.5
VALHALLA	**39.4 ± 0.3**	**31.7 ± 0.2**	**27.9 ± 0.3**	**33.0 ± 0.3**	**60.5 ± 0.1**	**52.3 ± 0.7**	**43.1 ± 0.3**	**52.0 ± 0.4**

Table 7: **BLEU score with CLIP Hallucination**, evaluated with Transformer-Small models on Multi30K.

multimodal machine translation. Our research can have a positive impact on many real-world applications of neural machine translation involving a broad range of languages. It improves translation performance in both well- and under-resourced scenarios which is of great practical importance. Negative impacts of our research are difficult to predict, however, it shares many of the pitfalls associated with standard MT models such as dataset/social bias and susceptibility to adversarial attacks. While we believe that these issues should be mitigated, they are beyond the scope of this paper.

References

[1] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58, 2014. 1, 2

[2] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 3

[3] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 1, 2, 3

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 4

[5] Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994. 1

[6] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017. 2

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[8] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019. 3

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 3

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 3

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3

[12] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 2

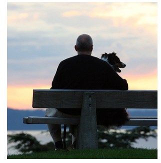
[13] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 1

[14] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*, 2021. 1, 2, 3

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[16] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 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*

Source EN A boy wearing a red shirt digs into the sand with a yellow shovel.
Reference DE Ein junge in einem roten shirt gräbt mit einer gelben schaufel im sand.
Text-Only Ein junge, der ein rotes hemd trägt, wirft einen gelben ball in den sand.
(*A boy wearing a red shirt throws a yellow ball in the sand.*)
VALHALLA Ein junge in einem roten hemd gräbt mit einer gelben schaufel in den sand.
(*A boy in a red shirt is digging in the sand with a yellow shovel.*)



Source EN A man sits on a bench holding his dog and looking at the water.
Reference DE Ein mann sitzt auf einer bank während er seinen hund hält und aufs wasser blickt.
Text-Only Ein mann sitzt auf einer bank und hält seinen hund.
(*A man is sitting on a bench and is holding his dog.*)
VALHALLA Ein mann sitzt auf einer bank und hält seinen hund und blickt auf das wasser.
(A man sits on a bench and holds his dog and looks out over the water.)

Source EN An overweight woman with long black hair in a pink shirt with a name tag is applying lipstick.
Reference FR Une femme en surpoids avec de longs cheveux noirs et un t-shirt rose avec un badge se met du rouge à lèvres.
Text-Only Une femme en surpoids avec de longs cheveux noirs est debout dans un micro.
(*An overweight woman with long dark hair is standing into a microphone.*)
VALHALLA Une femme en surpoids avec de longs cheveux noirs se maquille en chemise rose.
(*An overweight woman with long dark hair is applying makeup in pink shirt.*)

Source EN A child in blue and a child in white stand on a short concrete wall by a stream.
Reference FR Un enfant en bleu et un enfant en blanc sont debout sur un muret en béton au bord d'un ruisseau.
Text-Only Un enfant en t-shirt bleu et un enfant en béton.
(*A child in a blue t-shirt and a child in concrete.*)
VALHALLA Un enfant en bleu et un enfant en t-shirt blanc sont debout sur un mur de béton.
(*A child in blue and a child in a white t-shirt are standing on a concrete wall.*)

(a) Progressive Masking.

Source EN A child is riding a bike down an alleyway with graffiti in it.
Reference DE Ein kind fährt auf einem fahrrad durch eine gasse mit einem graffiti.
Text-Only Ein kind fährt mit seinem fahrrad eine straße entlang.
(*A child is riding his bicycle down a street.*)
VALHALLA Ein kind fährt auf einem fahrrad die straße entlang, die mit graffiti bemalt ist.
(*A child is riding a bicycle down the street that is painted with graffiti.*)

Source EN A skateboarder in a black t-shirt and jeans skating through the city.
Reference DE Ein skateboarder in einem schwarzen t-shirt und jeans fährt durch die stadt.
Text-Only Ein skateboarder in einem schwarzen t-shirt und mit sonnenbrille fährt mit drei personen die stadt entlang.
(*A skateboarder in a black t-shirt and with sunglasses drives three people along the city.*)
VALHALLA Ein skateboarder in einem schwarzen t-shirt und jeans fährt skateboard in der stadt.
(*A skateboarder in a black t-shirt and jeans is skateboarding in the city.*)

Source EN A guy in a white shirt is walking with a drink in his hand.
Reference FR Un gars en t-shirt blanc marche avec une boisson à la **main**.
Text-Only Un homme en t-shirt blanc marche avec une boisson dans sa **bouche**.
(*A man in a white t-shirt is walking with a drink in his* ***mouth***.)
VALHALLA Un homme en chemise blanche marche avec une boisson dans sa **main**.
(*A man in a white shirt is walking with a drink in his* ***hand***.)

Source EN The woman in blue is operating a camera in front of two other women.
Reference FR La femme en bleu manipule un appareil photo devant deux autres femmes.
Text-Only la femme en bleu manipule un rocher devant deux autres femmes.
(*The woman in blue is handling a rock in front of two other women.*)
VALHALLA La femme en bleu manipule une machine devant deux autres femmes.
(*The woman in blue operates a machine in front of two other women.*)

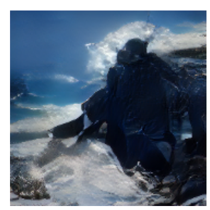
(b) Visual Entity Masking.

Figure 3: **Qualitative Translation Results with Progressive Masking and Visual Entiry Masking.** Phrases in gray in the source sentence are masked with <v> at model input. **VALHALLA** models generate more fluent and logical translations than text-only baseline transformer.

A surfer is doing a turn on his board.

A woman is performing gymnastics with long red ribbons.

Hockey player in white uniform with stick.



A younger woman sitting near a body of water with a dog.

Figure 4: **Reconstruction of Hallucinated Visual Tokens.** We use the pretrained VQGAN VAE image decoder to visualize the hallucinated visual sequence (the image decoder is *not* fine-tuned jointly with **VALHALLA**). **VALHALLA** captures abstract concepts such as "surfer" and "red ribbons", despite not being trained for high-quality image generation. Best viewed in color.

Joint Conference on Natural Language Processing, pages 6153–6166, 2021. 2, 3

[17] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[18] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2020. 3