VALHALLA: Visual Hallucination for Machine Translation

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

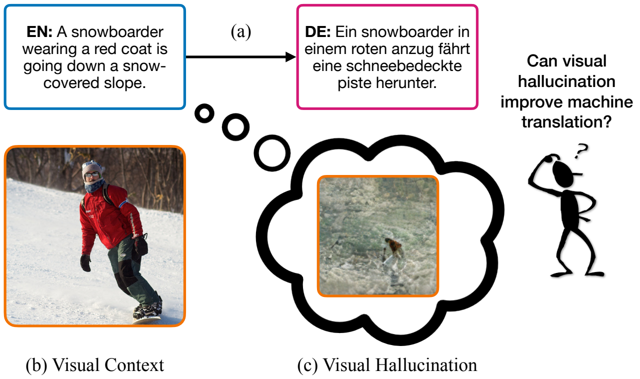
[1]UC San Diego, [2]MIT-IBM Watson AI Lab, [3]MIT CSAIL

Abstract

Designing better machine translation systems by considering auxiliary inputs such as images has attracted much attention in recent years. While existing methods show promising performance over the conventional text-only translation systems, they typically require paired text and image as input during inference, which limits their applicability to real-world scenarios. In this paper, we introduce a visual hallucination framework, called VALHALLA, which requires only source sentences at inference time and instead uses hallucinated visual representations for multimodal machine translation. In particular, given a source sentence an autoregressive hallucination transformer is used to predict a discrete visual representation from the input text, and the combined text and hallucinated representations are utilized to obtain the target translation. We train the hallucination transformer jointly with the translation transformer using standard backpropagation with cross-entropy losses while being guided by an additional loss that encourages consistency between predictions using either ground-truth or hallucinated visual representations. Extensive experiments on three standard translation datasets with a diverse set of language pairs demonstrate the effectiveness of our approach over both text-only baselines and state-of-the-art methods. Project page: http://www.svcl.ucsd.edu/projects/valhalla.

1. Introduction

Machine Translation (MT) is a core task in natural language processing and has undergone several paradigm shifts over the past few decades, from early rules-based systems [38] to pipelined statistical MT approaches [25, 33] to recent end-to-end neural network-based models [9, 58, 1, 62]. While such advances have led to impressive results on standard benchmarks, existing systems by and large utilize text-only information and lack any explicit grounding to the real world. There has thus been a growing interest in developing *multimodal* MT systems that can incorporate rich external information into the modeling process.


EN: A snowboarder wearing a red coat is going down a snow-covered slope.
(a)
DE: Ein snowboarder in einem roten anzug fährt eine schneebedeckte piste herunter.
Can visual hallucination improve machine translation?
?

(b) Visual Context (c) Visual Hallucination

Figure 1: Visual context such as images has been exploited in designing better machine translation systems. Different from most existing methods that require manually annotated sentence-image pairs as the input during inference, we introduce **VALHALLA**, that leverages hallucinated visual representation from the source sentences at test time for improved machine translation.

Consider the example in Figure 1(a), where a source sentence in English (blue box) is to be translated to a target sentence in German (red box). Since both sentences depict the same visual scene, shown in Figure 1(b), *there is common grounding information across the two sentences*. More generally, while there are many different ways to describe a situation in the physical world, the underlying visual perception is shared among speakers of different languages. The addition of visual context in the form of images is thus likely to help the machine translation. In particular, grounding should improve the data-efficiency of translation methods and benefit translation in low resource scenarios.

This has motivated much recent work on vision-based multimodal machine translation (MMT), which aims to improve machine translation systems by utilizing the visual modality [6, 30, 76, 20]. These methods typically require source sentences to be paired with the corresponding images during training *and* testing, which hinders their applicability

*Work done during an internship at the MIT-IBM Watson AI Lab.

to settings where images are not available during inference. In this work we consider the question of whether a system that has access to images only at training time can generalize to these settings. We hypothesize that *"visual hallucination, i.e., the ability to imagine visual scenes, can be leveraged to improve machine translation systems"*. Under this hypothesis, a translation system with access to images at training time could be taught to abstract an image or visual representation of the text sentence, as shown in Figure 1(c), in order to ground the translation process. At test time, this *abstracted* visual representation could be used in lieu of an actual image to perform multimodal translation.

We introduce a simple yet effective **V**isu**AL** **HALL**ucin**A**tion (**`VALHALLA`**) framework, which incorporates images at training time to produce a more effective text-only model for machine translation. As is usual for machine translation, the goal is to train a model that only sees source sentences at test time. However, during training, the model is trained to complement the text representation extracted from the source sentence with a latent visual representation that mirrors the one extracted from a real image (paired with the source sentence) by an MMT system. We achieve this by training an autoregressive hallucination transformer over a discrete codebook (learned using VQGAN-VAE [14]) to predict visual tokens from the input source sentences for multimodal translation.

`VALHALLA` consists of a pair of transformers: a visual hallucination transformer that maps the source sentence into a discrete image representation, and an MMT transformer that maps the source sentence paired with its discrete image representation into the target sentence. We train the transformer models end-to-end with a combination of hallucination, translation, and consistency losses. As sampling of the discrete image representations (i.e., visual hallucinations) is non-differentiable, we rely on a Gumbel-Softmax relaxation [21, 35] to effectively train the hallucination transformer jointly with the translation transformer. To the best of our knowledge, ours is the first work that successfully leverages an autoregressive image transformer jointly with the translation transformer to hallucinate discrete visual representations. We find that discrete visual representations lead to improved performance compared to continuous visual embeddings used in existing MMT methods [66, 30, 68, 74, 32].

Extensive experiments on three standard MT datasets (Multi30K [13], WIT [54] and WMT [2]) with a diverse set of language pairs and different scales of training data (in total 13 pairs) demonstrate the superiority of **`VALHALLA`** over strong translation baselines. **`VALHALLA`** yields an average 2$\sim$3 BLEU improvement over the text-only translation baseline, while consistently outperforming the most relevant state-of-the-art MMT methods that make use of continuous image representations [74, 32]. The gains over the text-only baseline are as large as $+3.1$ BLEU on under-resourced translation settings, such as the EN→RO and EN→AF tasks from WIT, confirming the hypothesis that visual hallucinations can have significant practical value in these settings. This is also confirmed by additional analysis suggesting that, under limited textual context, **`VALHALLA`** models do leverage visual hallucination to generate better translations.

2. Related Work

Multimodal Machine Translation. MMT has been studied from multiple perspectives [53, 64, 6, 76, 20, 69, 68, 31, 4]. Different from our work, a few methods [50, 57] use visual alignment for unsupervised word mapping and translation by retrieval. Unsupervised MMT methods have been proposed in [55, 19]. Recent works show that visual context does not help translation reliably [12, 66] or is mostly beneficial under limited textual context [5, 11]. Most MMT methods assume images as input at test time, which hinders their potential applications. Most relevant to our proposed approach are UVR-NMT [74] and ImagiT [32]. UVR-NMT uses a token-to-image lookup table to improve text-only NMT but requires retrieval of images during inference to match source language keywords. ImagiT uses a generative adversarial model to synthesize *continuous* image features for MMT. This differs from **`VALHALLA`**, which uses a hallucination model to predict *discrete* visual tokens from input text. In addition, ImagiT requires a computationally-heavy image captioning module, while our approach offers more flexible visual hallucination by using a transformer that autoregressively models text and image tokens as a single data stream.

Vision-Language Learning. Visual grounding has been used to improve performance and data-efficiency across many tasks [51, 37], such as semantic parsing [48], coreference resolution [26], representation learning [3, 23, 52], grammar induction [49, 75, 22, 18, 73], lexicon learning [63], and language learning with multimodal knowledge distillation [60], or mapping language tokens with images [59]. Conversely, image-text correspondence has also been exploited to improve vision tasks, such as image retrieval [41] and classification [44]. Despite recent progress, improving machine translation with no visual input at test time remains a challenging and largely under-addressed problem.

Text-to-Image Generation. Generating images from text has been extensively studied [45, 14, 47, 36]. Representative works use GANs [47, 67, 72, 43, 77, 71] to synthesize photo-realistic scenes with high semantic fidelity to their conditioned text descriptions. DALL-E [45] proposes an autoregressive transformer with discrete VAEs [39] to create images from text for a wide range of concepts expressible in natural language. While our approach is inspired by these works, the goal of the present work is to hallucinate discrete visual representations for improving machine translation instead of generating high-quality photo-realistic images.

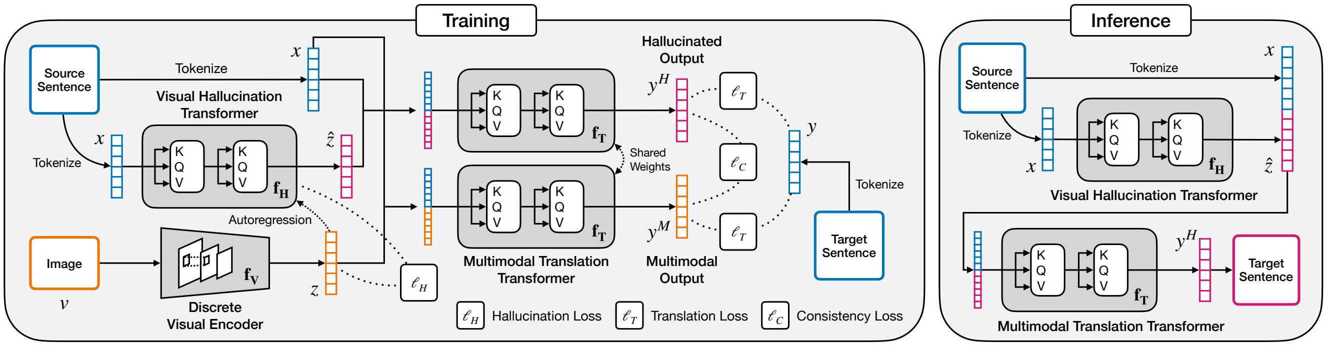


Figure 2: Overview of **VALHALLA** Architecture for Machine Translation. **Left**: Training pipeline of **VALHALLA**. Translation outputs are gathered from two streams of input, either with ground-truth visual tokens z or hallucinated representation $\hat{z}$, and optimized on a combination of *hallucination*, *translation* and *consistency* losses. **Right**: Inference process of **VALHALLA** in the absence of visual inputs.

Modality Hallucination. **VALHALLA** is also related to prior work on learning using side information [61]. A model to hallucinate depth features from RGB input for object detection is proposed in [17]. Graph distillation has been used to transfer multimodal privileged information across domains for action detection [34]. Garcia et al., [15] propose modality distillation for video action recognition.

3. Proposed Method

Given a corpus of source sentence $x \in \mathcal{X}$ and visual context $v \in \mathcal{V}$, typically images, our goal is to train a machine translation system that can translate a source sentence x into a sentence $y \in \mathcal{Y}$ in a target language without requiring images at inference time.

3.1. Preliminaries

Machine Translation. Contemporary MT systems are generally based on the encoder-decoder framework with attention [1, 62]. Given sequence pairs (x, y), where $x = (x_1, \ldots, x_S)$ is the source sentence of length S and $y = (y_1, \ldots, y_T)$ is the target sentence of length T, a transformer $\mathbf{f_T} = (\mathbf{f}_\mathbf{T}^{\text{enc}}, \mathbf{f}_\mathbf{T}^{\text{dec}})$ models the likelihood of target tokens conditioned on the input sequence as

$$\begin{aligned} p\left(y \mid x; \mathbf{f_T}\right) &= \prod_{i=1}^{T} \mathbf{f_T}\left(y_i \mid y_{<i}, x\right) \\ &\triangleq \prod_{i=1}^{T} \mathbf{f}_\mathbf{T}^{\text{dec}}\left(y_i \mid y_{<i}, \mathbf{f}_\mathbf{T}^{\text{enc}}(x)\right), \end{aligned} \tag{1}$$

where the decoder $\mathbf{f}_\mathbf{T}^{\text{dec}}$ predicts probability of output tokens at each location i by attending to encoder output $\mathbf{f}_\mathbf{T}^{\text{enc}}(x)$ and previous target tokens $y_{<i}$ using a cascade of attention layers. $\mathbf{f_T}$ is trained by minimizing the cross-entropy loss

$$\ell_T(\mathbf{f_T}) = \mathbb{E}_{(x,y)}\left[-\log p\left(y \mid x; \mathbf{f_T}\right)\right]. \tag{2}$$

Multimodal Machine Translation. MMT considers a visual input v as a complementary information source for machine translation. MMT systems typically use an encoder $\mathbf{f_V}$ to map an image into a latent visual representation $z = \mathbf{f_V}(v)$, which are fed into a modified decoder (e.g., by concatenating z with the word embeddings of x) to obtain the probabilities conditioned on visual input,

$$p\left(y \mid x, z; \mathbf{f_T}\right) = \prod_{i=1}^{T} \mathbf{f_T}\left(y_i \mid y_{<i}, x, z\right). \tag{3}$$

MMT models are trained on a dataset of triplets (x, v, y) by optimizing a translation loss based on cross-entropy

$$\ell_T(\mathbf{f_T}; z) = \mathbb{E}_{(x,z,y)}\left[-\log p\left(y \mid x, z; \mathbf{f_T}\right)\right]. \tag{4}$$

While incorporating visual information improves the translation performance of MMT systems over their text-only counterparts, it requires sentence-image pairs as input at *inference* time. This greatly limits the application of MMT systems in real world scenarios. We next introduce our **VALHALLA** framework, which addresses this constraint using discrete visual embedding and a hallucination module that predicts visual tokens from textual input for text-only translation.

3.2. Approach Overview

The overall **VALHALLA** framework is illustrated in Figure 2. The architecture consists of three neural network modules: A **discrete visual encoder** $\mathbf{f_V}$ for mapping input images into sequences of discrete tokens; a **hallucination transformer** $\mathbf{f_H}$ that predicts visual representations from the source sentence; and a **multimodal translation transformer** $\mathbf{f_T}$ that predicts the target sentence from the concatenated sequence of text and visual tokens.

During training, where input sentence-image pairs (x, v) are available, the translation output is predicted through two streams: *Multimodal* (bottom of Figure 2) and *hallucination* (top). The former uses ground-truth (discrete) visual representations z extracted from the input image, while the latter uses *hallucinated* representations $\hat{z}$. This produces two distributions y^M and y^H respectively, which are trained against the target sequence y with the cross entropy loss. Training losses also encourage consistency between predictions using either ground-truth or hallucinated visual representations, which is necessary for reliable performance of the visual hallucination module at inference time. As images associated with source sentences are not available at test time, the model utilizes the hallucination stream to generate pseudo-visual tokens and subsequently the translation output, conditioned on the unimodal text input x alone.

3.3. Discrete Visual Encoding

MMT is typically implemented by combining text input with *continuous* visual embeddings, such as convolutional features extracted from a pretrained ResNet [16]. In this work, we instead explore the use of a *discrete* visual encoder [39, 46, 45, 14]. This has two key advantages over a continuous embedding. First, images embedded into a sequence of discrete tokens can be easily concatenated with textual inputs (discrete word embeddings) into a multimodal sequence, which can then be processed by a single universal transformer to produce translation outputs. This vision-language fusion is nontrivial under continuous image representations, where complicated aggregation modules have been proposed for both MMT [66, 68, 30, 69] and other vision-language tasks [8, 65, 28]. Second, while regressing continuous visual representations requires careful design of losses and training schedule to prevent model predictions from collapsing to the mean value, visual hallucination in the discrete space reduces to a sequence-to-sequence learning problem trainable with a vanilla cross-entropy loss [45].

Motivated by this, we use discrete visual token sequences, which are essentially raster-scanned vector quantization maps of input images with respect to a feature codebook learned from training images. We implement vector quantization with the VQGAN VAE model of [14], using a visual encoder $\mathbf{f_V}$ to map input image v into a token sequence as

$$z = \mathcal{Q}(\mathbf{f_V}(v); E_V). \tag{5}$$

Here $z = [z_1, \dots, z_V]$ is a grid of discrete tokens laid out as a sequence where $z_i \in \{1, \dots, K\}$, $E_V = \{e^{(k)}\}_{k=1}^K$ are the d-dimensional visual codebook of size K, and $\mathcal{Q}$ denotes the quantization function

$$\mathcal{Q}_i(c; E_V) = \operatorname*{arg\,min}_{k \in \{1,\dots,K\}} \|c_i - e^{(k)}\|_2 \tag{6}$$

that maps each spatial location $i \in \{1, \dots, V\}$ of feature array $c = \mathbf{f_V}(v) \in \mathbb{R}^{V \times d}$ into the index of its closest visual code in E_V. Given a multimodal training set $\mathcal{D} = \{(x, v, y)\}$ where x, y denote source and target sentences, the image encoder $\mathbf{f_V}$ is trained on collection of images $\{v\}$ by optimizing a combination of reconstruction loss, vector quantization loss [39], and GAN adversarial loss. We refer the readers to [14] for more implementation details of the VQGAN VAE model.

Once $\mathbf{f_V}$ is learned, MMT feature aggregation becomes trivial as we can simply extend the input sequence of source tokens x to the translation transformer $\mathbf{f_T}$ with z encoded by (6) by concatenating the word/visual embeddings.

3.4. Visual Hallucination

During inference, when visual inputs are not available, **VALHALLA** relies on the visual hallucination module $\mathbf{f_H}$ to predict discrete visual tokens z given input text x. We follow [45] and implement an autoregressive transformer that models the concatenation of text and image tokens as

$$\begin{aligned} p(x, z; \mathbf{f_H}) &= p(x; \mathbf{f_H}) p(z \mid x; \mathbf{f_H}) \\ &= \prod_{i=1}^{S} \mathbf{f_H}(x_i \mid x_{<i}) \prod_{j=1}^{V} \mathbf{f_H}(z_j \mid z_{<j}, x). \end{aligned} \tag{7}$$

The hallucination transformer is trained to maximize the joint likelihood of x and z by optimizing the cross-entropy *hallucination loss*

$$\ell_H(\mathbf{f_H}) = \mathbb{E}_{(x,z)} \left[-\log p(x, z; \mathbf{f_H}) \right]. \tag{8}$$

We emphasize that as in [45] we model the joint $p(x, z; \mathbf{f_H})$ and not just the conditional $p(z \mid x; \mathbf{f_H})$, which was found to improve the results.

The hallucinated visual sequence $\hat{z}$ is then defined as the most likely token predicted by $\mathbf{f_H}$ at each time step i,

$$\hat{z}_i = \operatorname*{arg\,max}_{k \in \{1,\dots,K\}} \mathbf{f_H}(z_i = k \mid z_{<i}, x), \tag{9}$$

where the conditioning $z_{<i}$ is replaced with hallucinated visual sequence $\hat{z}_{<i}$ at inference time. While this enables the hallucination transformer to perform autoregressive decoding using source text tokens x only, it creates a mismatch between the training and inference, which is reflected in the output of the multimodal translation transformer. To reduce this mismatch, we define a *consistency loss*

$$\ell_C(\mathbf{f_H}, \mathbf{f_T}) = \mathbb{E}_{(x,z,y)} \left[\sum_{i=1}^{T} \mathrm{KL}[y_i^M \parallel y_i^H] \right], \tag{10}$$

where $y_i^M = p(y_i | x, z, y_{<i}; \mathbf{f_T})$ and $y_i^H = p(y_i | x, \hat{z}, y_{<i}; \mathbf{f_T})$ are the next word distributions from ground-truth visual tokens and hallucinated features respectively, and $\mathrm{KL}[y_i^M \| y_i^H]$ is the Kullback-Leibler divergence between the two conditional distributions.

3.5. Optimization

A remaining challenge for the joint optimization of the consistency loss of (10) and the translation loss of (4) is that the $\arg\max$ operator at the output of visual hallucination module (see (9)) prevents loss gradients from backpropagating through $\mathbf{f_H}$. To address this, we use the Gumbel-softmax relaxation [35, 21] during training, *i.e.*,

$$\hat{z}_i = \sum_{k=1}^{K} \frac{\exp((\log \pi_{i,k} + g_k)/\tau)}{\sum_l \exp((\log \pi_{i,l} + g_l)/\tau)} o_k, \tag{11}$$

where τ is the temperature of the softmax and o_k is a one-hot vector of length K activated at dimension k, $g_1, \ldots, g_K \sim \mathrm{Gumbel}(0, 1)$ i.i.d., and

$$\pi_{i,k} = \mathbf{f_H}(z_i = k \mid z_{<i}, x). \tag{12}$$

We set $\tau = 5$ as initial value and gradually anneal it down to 0 during training [21, 56], such that (11) converges to a one-hot distribution that resembles the use of (9) at inference.

The overall optimization objective of **VALHALLA** is finally defined as a weighted sum of translation loss, hallucination loss and consistency losses

$$\begin{aligned} \ell(\mathbf{f_H}, \mathbf{f_T}) = \ell_T(\mathbf{f_T}; z) + \ell_T(\mathbf{f_T}; \hat{z}) \\ + \gamma_H \ell_H(\mathbf{f_H}) + \lambda_C \ell_C(\mathbf{f_H}, \mathbf{f_T}), \end{aligned} \tag{13}$$

where γ_H is a hyperparameter that controls tradeoff between hallucination module $\mathbf{f_H}$ recovering ground-truth visual tokens ($\gamma_H \to \infty$) and extracting semantic information useful for machine translation ($\gamma_H \to 0$), and λ_C controls the degree of consistency between translation outputs.

Finally, we remark that our proposed approach can be seen as a version of latent variable MT where $z = [z_1, \ldots, z_V]$ are discrete latent variables that are grounded (i.e., imbued meaning) by being trained against "ground-truth" values of z obtained from the real images.

4. Experiments

4.1. Experimental Setup

Datasets and Tasks. We evaluate the performance of **VALHALLA** using three public datasets: Multi30K [13], Wikipedia Image Text (WIT) [54] and WMT2014 [2]. Multi30K [13] is a widely used MMT dataset, consisting of two multilingual expansions (DE and FR) of Flickr30K [70] dataset. We follow standard evaluation setup of [32, 66] to report performances on three test splits, Test2016, Test2017 and MSCOCO. WIT [54] is a large-scale multilingual dataset created by extracting text-image pairs from Wikipedia articles. As no prior work has studied MT on this dataset, we propose a new benchmark with seven language pairs under three settings, *well-resourced* (EN→{DE, ES, FR}), *under-resourced* (EN→RO, EN→AF), and *non-English* (DE→ES, ES→FR) splits. We use reference descriptions to obtain parallel sentence-image pairs. Detailed dataset preprocessing and cleaning procedure is provided in supplemental material.

WMT [2] is a widely-used text-only translation dataset, and we focus on the popular EN→DE and EN→FR tasks. We use the standard splits of WMT, and further construct two small sets created by sampling from the original set to investigate the performance of **VALHALLA** in under-resourced settings. Since WMT does not provide aligned images for training, we use CLIP [44] to retrieve top-5 images from Multi30K or WIT datasets to train our transformers.

Models. We experiment with different transformer model sizes (*Base*, *Small* and *Tiny*). Experiments on Multi30K use the *Small* and *Tiny* configurations, as smaller models have been shown to work better on this dataset [66]. For WIT and WMT tasks, we use the base configuration for the well-resourced tasks, while the small configuration is used for both the under-resourced and non-English tasks. See supplementary material for more detailed configurations.

Implementation Details. All our models are trained in three stages. First, we pretrain the discrete visual encoder $\mathbf{f_V}$ on the collection of images associated with training text; we then pretrain the hallucination transformer $\mathbf{f_H}$ using the loss of (8); finally, the translation transformer $\mathbf{f_T}$ is learned jointly with $\mathbf{f_H}$ on the combined loss of (13), with hyperparameters $\lambda_C = \gamma_H = 0.5$ determined by a grid search on validation data. Optimization is performed using Adam [24] with an inverse square root learning rate schedule and warm-up steps. During inference we use beam search with a beam size of 5.

Baselines. We compare with the following baselines. (1) text-only baseline that trains a transformer [62] without any visual information, (2) conventional MMT models (e.g., DCCN [30], GMNMT [69], and Gated Fusion [66]) that rely on sentence-image pairs for inference, (3) exiting methods where only text inputs are provided at test time for translation, including ImagiT [32], UVR-NMT [74], and RMMT [66]. We directly quote numbers reported in published papers when possible and use publicly available codes for UVR-NMT and RMMT on both WIT and WMT datasets.

Evaluation Metrics. We compute BLEU [40] and METEOR [10] scores to measure the translation performance of different models. Unless otherwise noted, we select the checkpoint with lowest validation loss for inference and further average the last ten checkpoints as in [66, 62], to compare with Gated Fusion/RMMT on Multi30K dataset.

4.2. Results on Multi30K

Table 1 shows the results on Multi30K. Transformer-Tiny (~ 20 times smaller than Transformer-Base) obtains the best performance in text-only translation, which is consistent with the recent findings in [66]. **VALHALLA** (denoted by $\mathbb{V}$ in Table 1) significantly outperforms the text-only baselines on all three test sets, which demonstrates the effectiveness of visual hallucination for text-only NMT. Using Transformer-Tiny as the backbone, **VALHALLA** obtains an average 35.4 BLEU in EN→DE and 54.4 BLEU in EN→FR, which is about 2.1 and 1.4 BLEU improvements over the text-only baseline.

Method	Model	EN → DE				EN → FR			
		Test2016	Test2017	MSCOCO	Average	Test2016	Test2017	MSCOCO	Average
Transformer-Base	T	32.0 ± 0.9	23.3 ± 0.8	21.3 ± 0.9	25.5 ± 0.9	59.7 ± 0.2	52.1 ± 0.1	42.4 ± 0.6	51.4 ± 0.3
Transformer-Small	T	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
	V	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
	VM	**39.6 ± 0.3**	**31.8 ± 0.2**	**27.9 ± 0.3**	**33.1 ± 0.3**	**60.5 ± 0.2**	**52.4 ± 0.6**	**43.4 ± 0.2**	**52.1 ± 0.3**
Transformer-Tiny	T	39.7 ± 0.3	31.7 ± 0.5	28.4 ± 0.2	33.3 ± 0.3	60.9 ± 0.5	53.7 ± 0.4	44.4 ± 0.2	53.0 ± 0.4
	V	**41.9 ± 0.2**	**34.0 ± 0.3**	30.3 ± 0.3	**35.4 ± 0.3**	62.3 ± 0.2	**55.1 ± 0.3**	**45.7 ± 0.2**	**54.4 ± 0.2**
	VM	**41.9 ± 0.2**	**34.0 ± 0.3**	**30.4 ± 0.4**	**35.4 ± 0.3**	**62.4 ± 0.3**	55.0 ± 0.3	**45.7 ± 0.4**	**54.4 ± 0.3**

Table 1: **BLEU score on Multi30K**. T: Baseline text-only transformer; V: **VALHALLA** model with hallucinated visual representations; VM: **VALHALLA** model with ground-truth visual representations. Please refer to supplementary material for METEOR score comparisons.

Method	EN → DE						EN → FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multimodal Machine Translation												
Gumbel-Attention [31]	39.2	57.8	31.4	51.2	26.9	46.0	–	–	–	–	–	–
CAP-ALL [29]	39.6	57.5	33.0	52.2	27.6	46.4	60.1	74.3	52.8	68.6	44.3	62.6
GMNMT [69]	39.8	57.6	32.2	51.9	28.7	47.6	60.9	74.9	53.9	69.3	–	–
DCCN [30]	39.7	56.8	31.0	49.9	26.7	45.7	61.2	76.4	54.3	70.3	45.4	65.0
VALHALLA (M)	**41.9**	**68.7**	**34.0**	**62.5**	**30.4**	**57.2**	**62.4**	**81.4**	**55.0**	**76.4**	**45.7**	**71.0**
Gated Fusion [66]	42.0	67.8	33.6	61.9	29.0	56.1	61.7	81.0	54.8	76.3	44.9	70.5
VALHALLA (M)	**42.6**	**69.3**	**35.1**	**62.8**	**30.7**	**57.6**	**63.1**	**81.8**	**56.0**	**77.1**	**46.4**	**71.3**
Text-Only Machine Translation												
VMMT$_F$ [7]	37.7	56.0	30.1	49.9	25.5	44.8	–	–	–	–	–	–
UVR-NMT [74]	36.9	–	28.6	–	–	–	58.3	–	48.7	–	–	–
ImagiT [32]	38.5	55.7	32.1	52.4	28.7	48.8	59.7	74.0	52.4	68.3	45.3	65.0
VALHALLA	**41.9**	**68.8**	**34.0**	**62.5**	**30.3**	**57.0**	**62.3**	**81.4**	**55.1**	**76.4**	**45.7**	**70.9**
RMMT [66]	41.4	68.0	32.9	61.7	30.0	56.3	62.1	81.3	54.4	76.1	44.5	70.2
VALHALLA	**42.7**	**69.3**	**35.1**	**62.8**	**30.7**	**57.5**	**63.1**	**81.8**	**56.0**	**77.1**	**46.5**	**71.4**

Table 2: **Comparison with state-of-the-art multimodal and text-only translation methods on Multi30K**. **VALHALLA** hallucinates visual representations from text-only inputs, while **VALHALLA (M)** uses ground-truth visual tokens at test time. Results in gray are computed with model averaging over 10 latest checkpoints. **VALHALLA** establishes new state-of-the-art for machine translation on Multi30K.

Moreover, **VALHALLA** has very similar performance with either hallucinated (V) or ground-truth representation (VM), showing strong ability to generate visual representations that are semantically consistent with the ground-truth.

Table 2 shows that **VALHALLA** outperforms all compared methods, achieving best BLEU and METEOR scores under both mulitmodal and text-only translation settings. While comparing to ImagiT [32], that generates continuous hallucinations via adversarial learning, **VALHALLA** obtains 2.3 and 1.9 BLEU improvements on the EN→DE and EN→FR tasks respectively, showing the effectiveness of discrete visual representations. Similarly, **VALHALLA** significantly outperforms UVR-NMT [74] in both tasks, without relying on additional image retrieval at test time. In summary, these consistent improvements clearly show that **VALHALLA** can effectively leverage visual semantics available at training time to greatly improve text-only test time translation.

We further divide the Test2016 set into different groups based on lengths of source sentences and compare performance with a text-only baseline in each group, as shown in Figure 3. **VALHALLA** consistently achieves the best performance in all groups, which once again confirms the effectiveness and generality of our approach. We further observe that the improvements are particularly pronounced for long sentences on both EN→DE and EN→FR tasks.

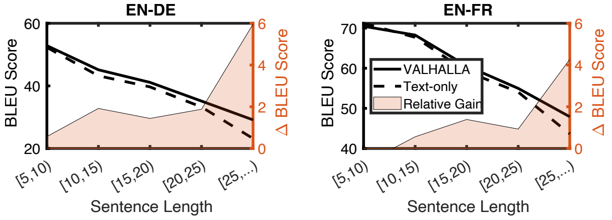


Figure 3: BLEU scores on different groups divided according to source sentence lengths on Multi30K Test2016 split.

4.3. Results on WIT

Table 3 shows that on WIT, **VALHALLA** again outperforms existing methods, improving text-only baseline performance from 15.1 to 17.2 BLEU, (see supplemental for METEOR scores). In particular, our approach achieves a substantial improvement over text-only baseline in under-resourced settings (2.9 on EN→RO and 3.2 on EN→AF). This shows that **VALHALLA** is more robust to conditions where the training corpora is small, revealing an important advantage of grounding information provided by visual hallucination for machine translation. Interestingly, while our approach is overall effective in translation between non-English languages, the improvement over text-only baseline is marginal. This is potentially due to an English-centric bias in the image-text pairs of original WIT dataset, which might

Method	Well-Resourced			Non-English		Under-Resourced		Average
	EN → DE	EN → ES	EN → FR	DE → ES	ES → FR	EN → RO	EN → AF	
Text-Only	16.0 ± 0.5	24.8 ± 0.8	16.1 ± 1.2	10.7 ± 0.2	16.2 ± 0.3	11.5 ± 0.7	10.8 ± 0.6	15.1 ± 0.6
UVR-NMT [74]	16.9 ± 0.2	26.4 ± 0.4	17.7 ± 0.3	10.9 ± 0.9	16.4 ± 0.6	12.5 ± 0.5	11.6 ± 1.7	16.1 ± 0.7
RMMT [66]	16.4 ± 0.3	24.8 ± 0.4	17.2 ± 1.6	11.0 ± 0.3	15.9 ± 0.7	9.9 ± 1.4	9.8 ± 1.0	15.0 ± 0.7
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**
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**

Table 3: **BLEU score on WIT**. Please refer to supplementary material for METEOR score comparisons.

Method	Visual Data	Well-Resourced				Under-Resourced			
		EN → DE		EN → FR		EN → DE		EN → FR	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Text-Only	–	27.1 ± 0.2	55.0 ± 0.1	39.1 ± 0.2	64.4 ± 0.1	16.7 ± 0.2	43.6 ± 0.2	25.9 ± 0.1	52.3 ± 0.3
UVR-NMT [74]	Multi30K	27.2 ± 0.2 (28.1)	55.3 ± 0.1	39.7 ± 0.2 (39.6)	64.9 ± 0.1	17.1 ± 0.1	44.1 ± 0.3	26.1 ± 0.3	52.8 ± 0.3
RMMT [66]		24.5 ± 0.2	52.8 ± 0.1	35.3 ± 0.0	61.2 ± 0.1	15.7 ± 0.2	41.9 ± 0.4	24.2 ± 0.3	50.7 ± 0.3
VALHALLA	Multi30K	**28.0 ± 0.1**	56.0 ± 0.1	**40.0 ± 0.1**	**65.2 ± 0.1**	17.6 ± 0.1	**44.8 ± 0.1**	**26.9 ± 0.2**	53.2 ± 0.2
	WIT	**28.0 ± 0.1**	**56.1 ± 0.1**	39.9 ± 0.1	65.1 ± 0.1	**17.7 ± 0.2**	44.7 ± 0.1	26.8 ± 0.0	**53.3 ± 0.1**
VALHALLA (M)	Multi30K	**28.0 ± 0.0**	56.0 ± 0.1	39.9 ± 0.1	65.0 ± 0.1	**17.7 ± 0.1**	**44.8 ± 0.2**	**26.9 ± 0.2**	**53.3 ± 0.3**
	WIT	27.9 ± 0.1	56.0 ± 0.2	39.8 ± 0.2	65.0 ± 0.1	**17.7 ± 0.2**	**44.8 ± 0.1**	26.8 ± 0.1	**53.3 ± 0.1**

Table 4: **Results on WMT2014**. UVR-NMT results in brackets are reported by the original paper.

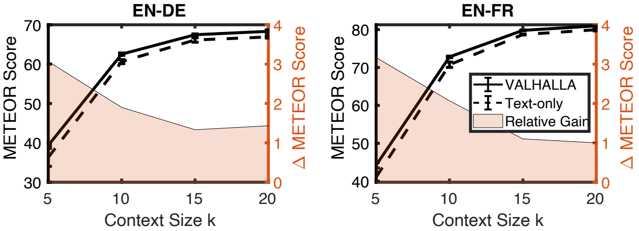


Figure 4: **Evaluation with Progressive Masking.** All results use METEOR scores on Multi30K Test2016 split.

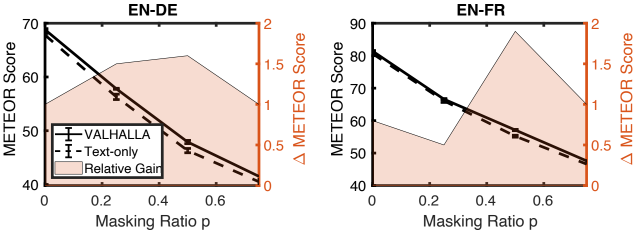


Figure 5: **Evaluation with Entity Masking.** All results use METEOR scores on Multi30K Test2016 split.

mean that visual modality fails to provide much additional information for translation between non-English languages.

4.4. Results on WMT

Table 4 shows the results on WMT. **VALHALLA** benefits from visual hallucination and outperforms all the compared methods in both well- and under-resourced settings. The improvements over text-only baseline are more significant in under-resourced scenarios, which is of significant practical value. We find that use of larger datasets, e.g., WIT instead of Multi30K for retrieving images at training time does not lead to substantial gain in performance, which is consistent with the previous findings [74]. Overall, the results on WMT show that our approach can be integrated into large-scale text-only translation datasets representing a wide variety of abstract concepts and real world entities (i.e., not specifically designed for multimodal machine translation).

4.5. Translation Under Limited Textual Context

We further study the robustness of **VALHALLA** framework for machine translation under limited textual context by degrading the input language modality during training and inference in two ways [5]: (1) Progressive masking that replaces all but the first k words of source sentences with a special token <v>, (2) Visual entity masking that randomly replaces visually grounded phrases (annotation from [42]) with probability p in the source sentence with <v>.

Source EN	Two children, a boy in a yellow shirt and a girl in blue and white stripes, swinging.
Reference FR	Deux enfants, un garçon en t-shirt jaune et une fille en rayures bleues et blanches, faisant de la balançoire.
Text-Only	Deux enfants, un garçon en t-shirt rayé bleu et une fille. (*Two children, a boy in a blue striped t-shirt and a girl.*)
VALHALLA	Deux enfants, un garçon en t-shirt jaune et une fille en t-shirt blanc, se balancent. (*Two children, a boy in a yellow t-shirt and a girl in a white t-shirt, are swinging.*)

Figure 6: **Qualitative Result with Progressive Masking.** Phrases in gray in the source sentence are masked with <v> at model input.

Progressive Masking. Figure 4 compares METEOR score of text-only baseline and **VALHALLA** as a function of context length k. On both EN→DE and EN→FR tasks, **VALHALLA** consistently outperforms the baseline under all settings. The gap between both methods widens as context size is reduced, with **VALHALLA** performing $\sim$ 3 METEOR points better. This suggests that visual hallucination is even more effective for translating ambiguous sentences out of context.

Visual Entity Masking. Figure 5 compares **VALHALLA** with text-only baseline when visual entities from the input source sentences are masked with probability p. Again, **VALHALLA** beats the text-only baseline in all test cases, with greatest improvements observed at $p = 0.5$. We attribute this to the effect of hallucination transformer inherently mod-

Backbone	Discrete Embedding	External Pretraining	Aggregation	EN-DE	EN-FR
CLIP RN-50	✗	CLIP	Gating	38.0	58.8
ResNet-50	✗	ImageNet	Gating	38.8	59.1
			Concatenation	38.3	60.0
VQGAN VAE	✓	None	Concatenation	**39.6**	**60.5**

(a) Discrete and continuous visual encoder backbones, evaluated with Transformer-Small on Multi30K Test2016 split.

Encoder Layers	Visual Token Length	EN-DE	EN-FR
4	$16^2 = 256$	13.5 ± 7.2	54.3 ± 0.4
5	$8^2 = 64$	36.3 ± 0.2	60.3 ± 0.2
6	$4^2 = 16$	**39.6 ± 0.3**	**60.5 ± 0.1**

(b) Visual encoder depths, evaluated with Transformer-Small on Multi30K Test2016.

Visual Data	Image Retrieval	EN-DE	EN-FR
Multi30K	✗	16.5 ± 0.3	26.2 ± 0.1
	✓	**17.6 ± 0.1**	**26.9 ± 0.2**
WIT	✗	16.6 ± 0.2	26.1 ± 0.3
	✓	**17.7 ± 0.2**	**26.8 ± 0.0**

(c) Training on WMT under-resourced tasks *without* image retrieval.

Table 5: **Ablation Studies.** All results use BLEU scores.

eling co-occurence between visual entities (e.g. human and objects) in the scene. This advantage reduces as masking ratio is further increased to 0.75, likely due to inability of visual hallucination to generate plausible predictions when majority of visual concepts are missing from input sentences.

Qualitative Examples. Figure 6 shows sample translation outputs from **VALHALLA** and text-only baseline under progressive masking, where **VALHALLA** successfully predicts masked phrase "swinging" through visual hallucination.

4.6. Ablation Analysis

Discrete vs. Continuous Representations. Table 5a compares performance of discrete VQGAN VAE with continuous representations, e.g., ResNet-50 [16] pretrained on ImageNet [27] or CLIP [44]. For ResNet, we consider two strategies to fuse multimodal inputs: *Gating* [69, 74, 66] learns a gating layer between text embeddings and pooled ResNet features; *Concatenation* flattens the feature map at conv5 block before global pooling, projects to the dimension of text embeddings and concatenates them, similar to the strategy used by **VALHALLA** with discrete tokens. Table 5a shows that ImageNet pretraining remains an effective way to extract continuous visual features, outperforming CLIP pretraining under the gating strategy. While aggregation by concatenation has no clear benefit over gating for continuous representations, it produces the strongest results for a discrete visual encoder. Importantly, since this strategy avoids pretraining encoders on large external datasets, it is potentially generalizable to a wider range of applications.

Visual Encoder Design. Table 5b shows the effect of varying depth of visual encoder, resulting in different lengths of encoded visual tokens. A smaller visual sequence is beneficial for multimodal modeling, as too many visual tokens may prevent the translation transformer from attending to the relevant text sequence and overfit to image inputs instead.

Joint Optimization. Compared to the jointly trained **VALHALLA** model, using a pretrained visual hallucination module off-the-shelf yields worse results (EN→DE BLEU: 39.0 vs 39.6 and 31.0 vs 31.7 with Transformer-Small model on Multi30K Test2016 and Test2017 respectively). This validates the necessity of jointly fine-tuning the hallucination transformer $\mathbf{f_H}$ and translation transformer $\mathbf{f_T}$ with the Gumbel-softmax sampling strategy outlined in (11).

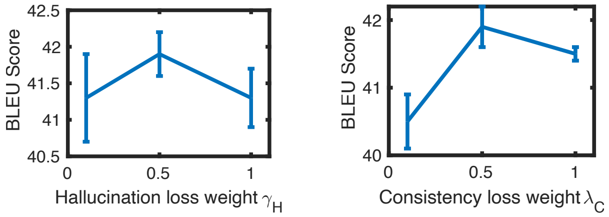


Figure 7: Influence of loss weights γ_H and λ_C of (13) on translation performance, measured on Multi30K EN→DE task.

Randomized Visual Tokens. On Multi30K'16 EN→FR, replacing inputs to MMT transformer $\mathbf{f_T}$ with *random* visual tokens reduced BLEU score to 61.2 from 62.3. We observe a similar drop (∼1–2 BLEU) in performance while using random visual tokens on other tasks as well, which suggests that hallucinated visual tokens are indeed of crucial significance.

Loss Hyperparameters. Figure 7 shows that **VALHALLA** is robust to the choice of hallucination weight γ_H but more sensitive to the consistency hyperparameter λ_C (1.4 BLEU improvement when increasing λ_C from 0.1 to 0.5). This shows that it is crucial to enforce consistency between translation *outputs* based on ground-truth and hallucinated features (10), in addition to consistency (8) in *visual* latent space.

Image Retrieval. We study the importance of image retrieval in training with the text-only corpora of WMT. Table 5c shows that the performance of **VALHALLA** trained with translation loss $\ell_T(\mathbf{f_T}; \hat{z})$ alone (i.e., directly using the hallucination transformer trained on Multi30K or WIT without retrieved images v) is considerably worse. This shows that the retrieved real images serve as important regularizer for the hallucination and translation transformers.

5. Conclusion

We present a new framework for improved machine translation by leveraging visual hallucination at test time, as opposed to existing MMT approaches based on sentence-image pairs. We utilize an autoregressive hallucination transformer to generate discrete visual representations from the input text and train it jointly with a multimodal translation transformer. We demonstrate effectiveness of our approach on three datasets, outperforming several competing methods.

Acknowledgements. This work was funded in part by NSF awards IIS-1924937 and IIS-2041009.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1, 3

[2] 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. 2, 5

[3] Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Incorporating visual semantics into sentence representations within a grounded space. *arXiv preprint arXiv:2002.02734*, 2020. 2

[4] Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Swaroop Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, 2021. 2

[5] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*, 2019. 2, 7

[6] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*, 2017. 1, 2

[7] Iacer Calixto, Miguel Rios, and Wilker Aziz. Latent variable model for multi-modal translation. *arXiv preprint arXiv:1811.00357*, 2018. 6

[8] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021. 4

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of EMNLP*, Oct. 2014. 1

[10] 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. 5

[11] Desmond Elliott. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, 2018. 2

[12] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177*, 2017. 2

[13] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 2, 5

[14] 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. 2, 4

[15] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 8

[17] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016. 3

[18] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. *arXiv preprint arXiv:2103.12975*, 2021. 2

[19] Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. Unsupervised multimodal neural machine translation with pseudo visual pivoting. *arXiv preprint arXiv:2005.03119*, 2020. 2

[20] Julia Ive, Pranava Madhyastha, and Lucia Specia. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*, 2019. 1, 2

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2, 5

[22] Lifeng Jin and William Schuler. Grounded PCFG Induction with Images. In *Proceedings of AACL*. Association for Computational Linguistics, 2020. 2

[23] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*, 2017. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. 1

[26] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014. 2

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 8

[28] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4654–4662, 2019. 4

[29] Zhifeng Li, Yu Hong, Yuchen Pan, Jian Tang, Jianmin Yao, and Guodong Zhou. Feature-level incongruence reduction for multimodal translation. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 1–10, 2021. 6

[30] Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329, 2020. 1, 2, 4, 5, 6

[31] Pengbo Liu, Hailong Cao, and Tiejun Zhao. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*, 2021. 2, 6

[32] Quanyu Long, Mingxuan Wang, and Lei Li. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, 2021. 2, 5, 6

[33] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49, 2008. 1

[34] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018. 3

[35] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 2, 5

[36] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017. 2

[37] Tobias Norlund, Lovisa Hagström, and Richard Johansson. Transferring knowledge from vision to language: How to achieve it and how to measure it? *arXiv preprint arXiv:2109.11321*, 2021. 2

[38] Eric H. Nyberg III and Teruko Mitamura. The KANT system: Fast, accurate, high-quality translation in practical domains. In *COLING 1992 Volume 3: The 14th International Conference on Computational Linguistics*, 1992. 1

[39] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 2, 4

[40] 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. 5

[41] Jose Costa Pereira and Nuno Vasconcelos. Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems. *Computer Vision and Image Understanding*, 124:123–135, 2014. 2

[42] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 7

[43] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019. 2

[44] 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. 2, 5, 8

[45] 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, 4

[46] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019. 4

[47] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 2

[48] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. *arXiv preprint arXiv:1906.02890*, 2019. 2

[49] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually Grounded Neural Syntax Acquisition. In *Proceedings of ACL*, 2019. 2

[50] Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10850–10859, 2020. 2

[51] Damien Sileo. Visual grounding strategies for text-only natural language processing. *arXiv preprint arXiv:2103.13942*, 2021. 2

[52] Karan Singhal, Karthik Raman, and Balder ten Cate. Learning multilingual word embeddings using image-text data. *arXiv preprint arXiv:1905.12260*, 2019. 2

[53] Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 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, 2016. 2

[54] 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. 2, 5

[55] Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2019. 2

[56] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. *arXiv preprint arXiv:1911.12423*, 2019. 5

[57] Dídac Surís, Dave Epstein, and Carl Vondrick. Globetrotter: Unsupervised multilingual translation from visual alignment. *arXiv preprint arXiv:2012.04631*, 2020. 2

[58] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NeurIPS*, 2014. 1

[59] Hao Tan and Mohit Bansal. Vokenization: improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*, 2020. 2

[60] Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *arXiv preprint arXiv:2107.02681*, 2021. 2

[61] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 3

[62] 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. 1, 3, 5

[63] Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie Francine Moens. Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 188–194, 2016. 2

[64] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 2

[65] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 4

[66] 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 Joint Conference on Natural Language Processing*, pages 6153–6166, 2021. 2, 4, 5, 6, 7, 8

[67] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[68] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, 2020. 2, 4

[69] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. *arXiv preprint arXiv:2007.08742*, 2020. 2, 4, 5, 6, 8

[70] 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. 5

[71] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021. 2

[72] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[73] Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. Video-aided Unsupervised Grammar Induction. In *Proceedings of NAACL*, 2021. 2

[74] 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. 2, 5, 6, 7, 8

[75] Yanpeng Zhao and Ivan Titov. Visually Grounded Compound PCFGs. In *Proceedings of EMNLP*, 2020. 2

[76] Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. *arXiv preprint arXiv:1808.08266*, 2018. 1, 2

[77] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 2