40th AAAI Conference on Artificial Intelligence (AAAI), Singapore, 2026.

Scaling Laws for Conditional Emergence of Multilingual Image Captioning via Generalization from Translation

Julian Spravil[1,3*], **Sebastian Houben**[2,1], **Sven Behnke**[3,4,5,1]

[1]Fraunhofer IAIS, Germany
[2]Institute for Artificial Intelligence and Autonomous Systems, University of Applied Sciences Bonn-Rhein-Sieg, Germany
[3]Autonomous Intelligent Systems, Computer Science Institute VI, University of Bonn, Germany
[4]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
[5]Center for Robotics, University of Bonn, Germany

Abstract

Cross-lingual, cross-task transfer is challenged by task-specific data scarcity, which becomes more severe as language support grows and is further amplified in vision-language models (VLMs). We investigate multilingual generalization in encoder-decoder transformer VLMs to enable zero-shot image captioning in languages encountered only in the translation task. In this setting, the encoder must learn to generate generalizable, task-aware latent vision representations to instruct the decoder via inserted cross-attention layers. To analyze scaling behavior, we train Florence-2 based and Gemma-2 based models (0.4B to 11.2B parameters) on a synthetic dataset using varying compute budgets. While all languages in the dataset have image-aligned translations, only a subset of them include image captions. Notably, we show that captioning can emerge using a language prefix, even when this language only appears in the translation task. We find that indirect learning of unseen task-language pairs adheres to scaling laws that are governed by the multilinguality of the model, model size, and seen training samples. Finally, we demonstrate that the scaling laws extend to downstream tasks, achieving competitive performance through fine-tuning in multimodal machine translation (Multi30K, CoMMuTE), lexical disambiguation (CoMMuTE), and image captioning (Multi30K, XM3600, COCO Karpathy).

1 Introduction

Multilingual image-to-text modeling is a fundamental step towards achieving universal accessibility of multimedia content. Recent advancements in vision-language models (VLMs) demonstrate impressive results across various tasks such as image understanding and visual question answering (Liu et al. 2023; Xiao et al. 2024; Gemma Team 2025). This progress is driven by the availability of large, primarily English vision-language datasets. Two main approaches enable cross-lingual transfer, extending capabilities from English to other languages. The first method involves fine-tuning multilingual models in a single language for a specific task, while keeping the embeddings and most layers of the language model frozen in order to retain its multilingual representations. (Wu and Dredze 2019; Chen et al. 2023; Futeral et al. 2025a). Models such as mBERT (Devlin et al. 2019) and NLLB (Costa-jussà et al. 2022) are common choices, representing established multilingual large language models (LLMs) and machine translation models (MTMs), respectively. Evidence suggests that both multilingual and monolingual models learn language-agnostic representations, enabling cross-lingual transfer (Libovický, Rosa, and Fraser 2019; de Souza et al. 2024). The second method uses continuous pre-training on collected or generated multilingual data (Gogoulou et al. 2022; Qiu et al. 2022; Futeral et al. 2025b).

Current methods are fundamentally constrained by their underlying models and data. By using pre-trained LLMs, they inherit the trade-off between performance and language coverage (Conneau et al. 2020). Generating data requires ca-

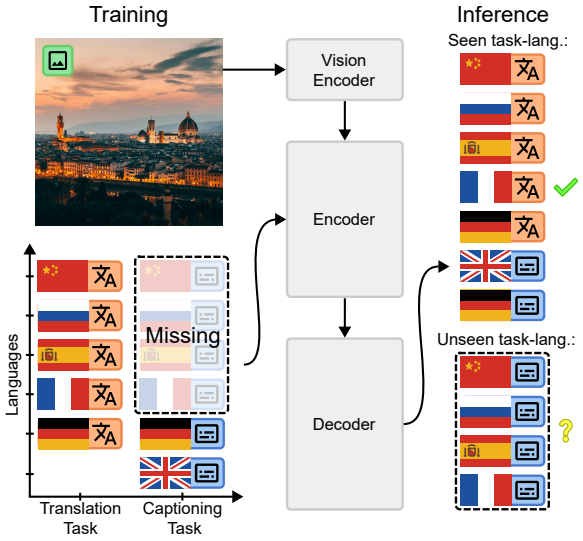

Figure 1: We train a vision-language model (VLM; middle) on an incomplete dataset (left) that covers the tasks image captioning (blue) and multimodal machine translation (orange). While En→X translation is available for all languages, captioning data is limited to only English and German. The VLM generalizes to the missing captioning-language pairs with sufficient scale (right).

*Corresponding author: julian.spravil @ iais.fraunhofer.de

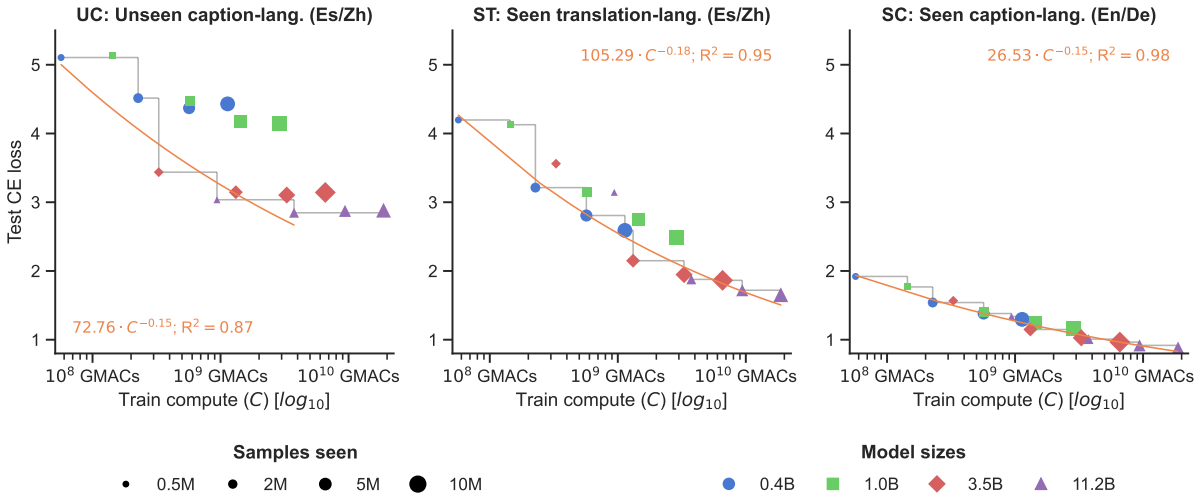

Figure 2: Test cross-entropy (CE) loss for various training compute budgets (GMACs, Giga multiply-accumulate operations). We show results for the test splits for unseen captioning (UC) in Spanish (Es) and Chinese (Zh), seen translation (ST) in the same languages, and seen captioning (SC) in English (En) and German (De). All models are trained for 0.5M, 2M, 5M, and 10M seen samples. Equation 1 is fitted to the points on the Pareto frontier (gray staircase graph). Higher compute budgets improve CE loss for UC (left), ST (middle), and SC (right). This suggests that translation facilitates generalization in captioning.

pable MTMs while collecting sufficient data is impractical. Furthermore, multimodal approaches struggle to resolve lexical ambiguities (e.g., distinguishing between "bat" as an animal or sports equipment) (Futeral et al. 2023). The dynamics of multilingual cross-task generalization, particularly at scale, remains largely unexplored, despite its implications for the necessity of collecting data for each task in every language. To overcome these challenges, exploring systematic generalization (Fodor and Pylyshyn 1988; Lake and Baroni 2023) is a critical next step.

We explore the scaling laws of generalization within a realistic multimodal setting using a partially pre-trained encoder-decoder transformer (Vaswani et al. 2017) and a standard training method. Our goal is to learn a set of task and language combinations and transfer these capabilities to different task-language combinations in a zero-shot manner by scaling, as illustrated in Figure 1. In summary, our main contributions are:

- We investigate the scaling laws of model performance on seen task-language data and its generalization to unseen task-language data, analyzing the effects of model size, number of seen training samples, and initial cross-entropy loss. We show that generalization by only learning translation to facilitate captioning is influenced not only by multilingual pre-training but also by model scale and seen training samples.
- We find that the observed scaling trends persist during fine-tuning, resulting in competitive performance across multiple benchmarks (Multi30K, CoMMuTE, COCO Karpathy, and XM3600).
- We present a modular encoder-decoder framework built on pre-trained VLMs and LLMs, with sizes ranging from 0.4B to 11.2B parameters.
- We propose a pipeline that generates synthetic multilingual image captions and aligns a text-only translation dataset to these images using contrastive VLMs and off-the-shelf MTMs.

2 Related Work

Scaling laws. Following a power-law relationship, scaling laws enable predictable and efficient large-scale training and offer valuable insights into training dynamics. These laws were first described for computer vision (Sun et al. 2017) and have since been applied to natural language processing (Kaplan et al. 2020; Ghorbani et al. 2022; Hoffmann et al. 2022; Fernandes et al. 2023), transfer learning (Hernandez et al. 2021), and contrastive vision-language learning (Cherti et al. 2023). Current studies only focus on what is explicitly learned and not on what is implicitly learned.

Machine translation. The transformer (Vaswani et al. 2017) has revolutionized machine translation, with impressive results through pre-training on extensive translation data (Costa-jussà et al. 2022), non-parallel multilingual data (Devlin et al. 2019), and with weak supervision (Conneau et al. 2020). While models like NLLB (Costa-jussà et al. 2022) demonstrate significant success by supporting 200 languages, many translation directions remain under-resourced. Several techniques have been used to address this issue, including language pivots (Wu and Wang 2007), the generation of pseudo labels (Firat et al. 2016), and leveraging similar languages or parallel data (Johnson et al. 2017).

Cross-lingual transfer. Multilingual LLMs (Devlin et al. 2019; Liu et al. 2020) and MTMs (Costa-jussà et al. 2022) exhibit strong cross-lingual transfer performance, even when fine-tuned with monolingual data (Wu and Dredze 2019; Pires, Schlinger, and Garrette 2019; Muennighoff et al. 2023). Studies show that language-neutral and

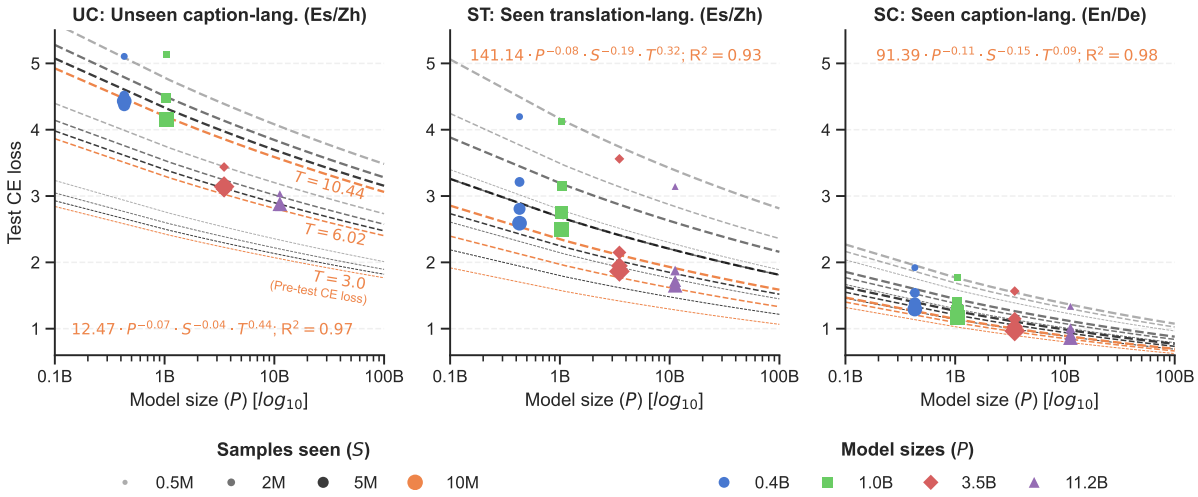

Figure 3: Test CE loss as a function of model size (P), number of seen samples (S), and initial CE loss (T) across the three test splits: UC, ST, and SC. The dashed lines represent the fitted functions for three values of T: $T = 10.44$ for Florence-2 based models, $T = 6.02$ for Gemma-2 based models, and $T = 3.0$ for a hypothetical highly multilingual VLM. Line thickness is proportional to the T value. The measured results for all evaluated models are shown as points. The 10M seen-sample line is highlighted in orange, while lower sample counts are represented by progressively lighter shades of gray. Notably, for the UC task, test CE loss decreases as P and S increase and T decreases.

language-specific components develop and facilitate transfer (Libovický, Rosa, and Fraser 2019; de Souza et al. 2024).

Multimodal multilingual learning. The issue of data scarcity is addressed by adapting multilingual models with machine translated data (Futeral et al. 2025a), small multimodal multilingual datasets (Mitzalis et al. 2021; Futeral et al. 2023; Hirasawa et al. 2023), text-only translation data (Mitzalis et al. 2021; Hirasawa et al. 2023), and image captioning as auxiliary task (Mitzalis et al. 2021). Web-crawled multilingual vision data can unlock few-shot capabilities (Futeral et al. 2025b). Adapting monolingual models can also be effective, although the size of the pre-training corpus can impact their final performance, while language similarity has little effect (Gogoulou et al. 2022). Text-only models can solve some multimodal multilingual tasks that have few vision-dependent samples (Hirasawa et al. 2023; Futeral et al. 2023). Resolving lexical ambiguities requires additional context (Futeral et al. 2023). Contrastive models (Radford et al. 2021; Carlsson et al. 2022; Chen et al. 2023) excel at resolving ambiguities, whereas multilingual generative models struggle (Futeral et al. 2023, 2025a).

Muennighoff et al. (2023) also perform cross-lingual transfer by fine-tuning LLMs without target-task data but rely on the model's existing multilingual capabilities. In contrast, we investigate scaling laws for improving performance in a new task-language combination with respect to model size, seen training samples, and initial multilingual loss.

3 Models and Datasets

We construct partially pre-trained VLMs based on the pre-trained models Florence-2 (Xiao et al. 2024) and Gemma-2 (Gemma Team 2024). Florence-2 is an encoder-decoder VLM supporting tasks ranging from object detection to image captioning and is available with 0.2B and 0.8B parameters. The encoder generates a sequence of tokens, representing both image and task, that is used to instruct the decoder via cross-attention layers.

To obtain larger model sizes, we combine Gemma-2, an LLM with sizes of 3B and 9B parameters, with the encoder of Florence-2. The image-task encoder outputs are integrated into the decoder by inserted cross-attention layers that are weighted with a learnable parameter initialized with zero following Flamingo (Alayrac et al. 2022).

For the decoder, we reuse the tokenizer of Gemma-2 with a vocabulary size of 256k. For the two smaller models with the Florence-2 decoder, we reinitialize the embedding layer and language modeling head to fit the Gemma-2 tokenizer using the method by Gee et al. (2022). The encoder retains the original tokenizer and embeddings from Florence-2.

This leads to the standard transformer encoder-decoder VLM designed for continuous pre-training, available in sizes 0.4B, 1B, 3.5B, and 11.2B.

3.1 Continuous Pre-training Dataset

We create a synthetic training dataset tailored for this study based on CC12M (Changpinyo et al. 2021) and CCMatrix (Schwenk et al. 2021). The dataset covers six languages: English (En), German (De), French (Fr), Spanish (Es), Russian (Ru), and Chinese (Zh). CC12M contains 12M web-crawled images of which 10M are available. We pair the images with generated image descriptions sourced from Hugging Face[1] and translated to German with NLLB-3.3B (Costa-jussà et al. 2022). CCMatrix is a web-crawled corpus covering 38 languages with 6.8B parallel sentences

[1]hf.co/datasets/CaptionEmporium/conceptual-captions-cc12m-llavanext

Task	Coefficient	Estimate [95% CI]	p-value
SC	β_1	-0.59 [-0.79, -0.38]	$p < 0.001$
	β_2	-0.72 [-0.80, -0.63]	$p < 0.001$
	β_3	0.10 [-0.10, 0.31]	$p = 0.293$
ST	β_1	-0.36 [-0.74, 0.03]	$p = 0.068$
	β_2	-0.73 [-0.89, -0.57]	$p < 0.001$
	β_3	0.29 [-0.09, 0.68]	$p = 0.125$
UC	β_1	-0.41 [-0.69, -0.12]	$p = 0.009$
	β_2	-0.23 [-0.35, -0.11]	$p = 0.001$
	β_3	0.57 [0.29, 0.85]	$p < 0.001$

Table 1: Standardized coefficients for the second power law (Equation 2) for UC, ST, and SC in $\log_{10}$ space.

of which about 661M are aligned with English. We extract translations from English to the aforementioned target languages. Accelerated by Faiss (Douze et al. 2024), we use CLIP-ViT-B/16 (Radford et al. 2021) to align the sentences with the images of CC12M via top-5 matching and subsequent deduplication. See the appendix for more details.

In total, the training dataset contains 10M images aligned with 32M captions in En and De and 105M translation pairs for En→{De, Fr, Es, Ru, Zh}. To evaluate generalization, we intentionally omit captioning data for {Fr, Es, Ru, Zh}.

For the test set, we use a subset of 4.4K images from CC12M and create captioning data for two representative languages for the unseen task-language pairs (Es and Zh). We divide the test set into three parts: unseen captioning (UC) with 4.4K Es and 3.5K Zh captions, seen translation (ST) with 4.1K En→Es and 3.8K En→Zh translations, and seen captioning (SC) with 4.4K En and 3.1K De captions. Note that "seen" refers to the task-language combination being part of training, not the specific data instances.

3.2 Downstream Tasks Dataset

We construct a fine-tuning dataset that includes a mix of downstream tasks with full language coverage, starting with the train split of Multi30K (Elliott et al. 2016) for translation (Task 1, En→{De, Fr}). For captioning, we include Multi30K (Task 2) for short En and De captions, Image Paragraph (Krause et al. 2017) for detailed captions, and DOCCI (Onoe et al. 2024) for highly detailed descriptions. Missing languages are added to the aforementioned datasets with neural machine translation (Costa-jussà et al. 2022). Additionally, we include the train/restval split of COCO Karpathy (Chen et al. 2015; Karpathy and Fei-Fei 2017). In total, the fine-tuning dataset has 166K images with 1.6M captions of different styles and 145K translation samples covering all task-language combinations.

4 Scaling Laws

We explore the scaling laws of continuous pre-training in a multilingual multi-task scenario, where not all task-language combinations are given within the training data. The relationship between cross-entropy (CE) loss and training compute can be described by a power law, where changes in model size and seen samples result in predictable, non-linear improvements (Kaplan et al. 2020; Ghorbani et al. 2022; Hoffmann et al. 2022; Fernandes et al. 2023).

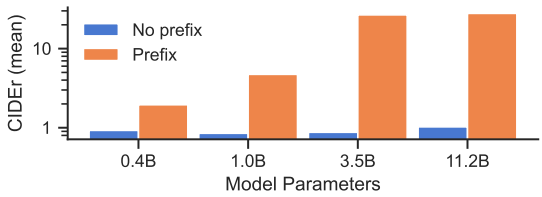

Figure 4: Effect of adding a prefix (Fr: "La photo montre", etc.) to the decoder input to unlock zero-shot captioning. Tested on the image captioning dataset XM3600 in the unseen languages Fr, Es, Ru, and Zh. The mean CIDEr over unseen languages significantly improves with the prefix.

Our first power law selects the Pareto frontier from all data points, an approach inspired by Cherti et al. (2023). The scaling law to predict the CE loss y from the training compute C and error term ϵ is given by:

$$y = \alpha_0 C^{\alpha_1} + \epsilon, \tag{1}$$

where α_0 and α_1 are the parameters to be estimated. The total computational cost C in multiply–accumulate operations (MACs) is estimated by $C = S \cdot F \cdot (1 + P_t/P)$, where S is the number of seen training samples, F is the forward pass MACs estimated with `fvcore`[2], P_t is the number of trainable parameters, and P is the total number of parameters.

The second, multivariate power law predicts the CE loss y based on the seen training samples S, the model parameters P, the mean initial CE loss T of the base model on the UC and ST test sets, and the error term ϵ:

$$y = \beta_0 P^{\beta_1} S^{\beta_2} T^{\beta_3} + \epsilon, \tag{2}$$

where β_0, β_1, β_2, and β_3 are to be estimated. In developing our model, we first considered a baseline using only variables S and P as commonly done in the literature. However, the omission of variable T led to a biased result. We transform Equations 1 and 2 into $\log_{10}$ space to enable a linear regression analysis with ordinary least squares (OLS).

Training setup. We train all our models using AdamW (Loshchilov and Hutter 2017) with CE loss, a batch size of 1024, a weight decay of 0.01, and a learning rate of $1e^{-4}$, which is scheduled with a linear warm-up for 100 steps and cosine decay. The input length for encoder and decoder is truncated to a maximum length of 128 and the image resolution is set to 224 px. Each model scale is trained for 500, 2K, 5K, and 10K iterations, where the latter corresponds to roughly one full epoch. We use online balancing to sample equally from each task-language combination. While the vision encoder is always frozen, we freeze the decoder layers of the 3.5B and 11.2B models as well. This means that the vision-task encoder, the inserted cross-attention layers, the language modeling head, and the embedding layer are trainable. Using the `transformers` library (Wolf et al. 2020), we trained on a HPC node equipped with 4 NVIDIA H100 GPUs.

[2] github.com/facebookresearch/fvcore

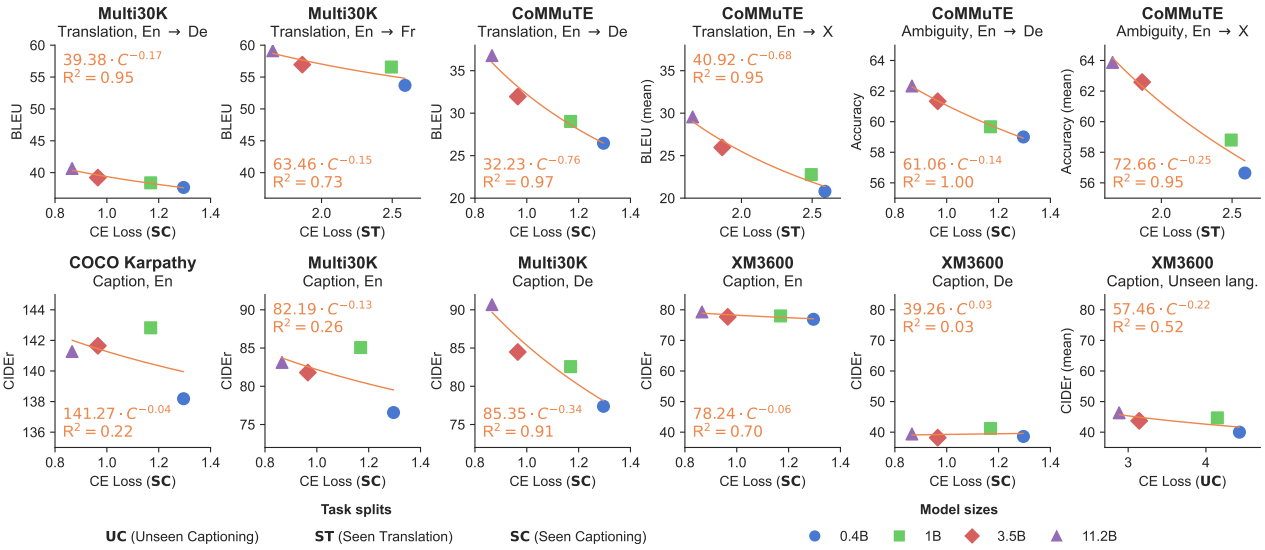

Figure 5: Downstream task performance with respect to CE loss, measured on the UC, ST, and SC tasks, depending on the type of downstream task. First row: Multi30K translation to De and Fr measured in BLEU (Task 1; mean over Test2016, Test2017 and AmbiguousCOCO splits), CoMMuTE translation and disambiguation for En→De and En→{De, Fr, Ru, Zh} measured in BLEU and accuracy, respectively. Second row: Captioning tasks measured with CIDEr: COCO Karpathy (En), Multi30K (En, De) (Task 2, Test2016), and XM3600 for En, De, and unseen languages (Fr, Es, Ru, Zh). We use a consistent y-axis scale for matching dataset and task.

Evaluation setup. We calculate the CE loss y over the three test sets of our continuous pre-training dataset covering the settings: unseen captioning (UC) in Es and Zh, seen translation (ST) in Es and Zh, and seen captioning (SC) in En and De. The initial CE loss T is calculated as a measure of multilinguality on the UC and ST test sets before training is conducted on the untrained but restructured models.

4.1 Results

The CE loss of the runs with compute budget and the fit of the first power law (Equation 1) are visualized in Figure 2. The analysis confirms a strong fit for SC (R^2=0.98) and ST (R^2=0.95) and a slightly weaker fit for UC (R^2=0.87). All models exhibit a clear inverse correlation between CE loss and train compute, which, while expected for seen tasks (SC and ST), suggests that models can generalize to unseen tasks (UC) in a language encountered only through translation.

Figure 3 presents the observed unstandardized coefficients of the second power law (Equation 2) along with extrapolations for larger and more multilingual models. To compare the relative importance of predictors across models, we additionally report standardized coefficients (standardized across the combined SC, ST, and UC tasks) in Table 1. All three models have a strong fit with R^2=0.98, R^2=0.93, and R^2=0.97 for SC, ST, and UC, respectively. Regression diagnostics indicate that OLS assumptions are satisfied.

The SC model is the standard setting with En and De captioning in the training and evaluation data. The standardized coefficients reveal that both training samples S (β_2=−0.72) and model size P (β_1= − 0.59) are negatively correlated with the test CE loss. In contrast, a lower initial CE loss T is weakly connected with a lower predicted loss (β_3=0.10), though this effect has high uncertainty. Though the negative dependencies on S and P are consistent with standard scaling laws, our findings contradict those of Zheng et al. (2024), who suggest that model size is more important than dataset size for the continuous pre-training of LLMs for cross-lingual transfer. We found no evidence that either predictor is more important, as the 95% confidence interval (CI) of the estimated difference between coefficients includes zero ($\beta_1-\beta_2$=0.13, 95% CI [-0.09, 0.35]). This suggests that known scaling behaviors may differ for multimodal tasks.

The ST model covers multilingual machine translation from En to Es and Zh. In this setting, only the number of training samples S (β_2= − 0.73) has a negative effect on test CE loss. The standardized coefficients for both model size P (β_1= − 0.36) and initial CE loss T (β_3=0.29) remain inaccurate. The lack of a clear effect is unexpected, as larger, more multilingual models are presumed to perform better. An explanation for this finding and the model's lower R^2 value is that machine translation may require separate terms for encoder and decoder parameters (Ghorbani et al. 2022). In our setup, the encoder is relatively small, which may introduce a bottleneck.

The UC model tests generalization to unseen image captioning in Es and Zh. Standardized coefficients show a positive association for initial CE loss T (β_3=0.57) and negative associations for model size P (β_1= − 0.41) and seen training samples S (β_2= − 0.23). By comparing the magnitudes, we identify that the influence of initial CE loss T is greater than that of training samples S ($\beta_3+\beta_2$=0.34, 95% CI [0.03, 0.65]). In contrast, the differences in magnitude between T and model size P ($\beta_3+\beta_1$=0.16, 95% CI [-0.39, 0.72]) and

	Multi30K Translation		CoMMuTE Translation		CoMMuTE Ambiguity		COCO Caption	Multi30K Caption		XM3600 Caption		
	En→De	En→Fr	En→De	En→X	En→De	En→X	En	En	De	En	De	Unseen
Model	BLEU	BLEU	BLEU	BLEU	Acc.	Acc.	CIDEr	CIDEr	CIDEr	CIDEr	CIDEr	CIDEr
Gemma-3-12B*	39.2	52.2	**44.1**	**38.5**	**73.3**	**76.6**	48.1	50.8	55.5	34.0	39.6	46.6
Pixtral-12B*	37.9	53.8	40.7	35.9	73.3	75.5	61.1	62.5	64.4	71.1	38.3	**50.5**
Baseline-6B*	37.3	54.3	41.5	32.5	61.7	61.1	**145.2**	84.0	50.4	**82.0**	38.2	45.1
Florence-2-L							143.3					
PaliGemma-3B*							141.7	**88.9**	57.6	79.1	37.7	48.5
MOF	24.9	35.1			63.7	66.5						
ZeroMMT-3.3B	37.1	53.3			60.8	62.2						
VGAMT	37.4	58.4			57.1							
NLLB-3.3B*	37.4	53.7	40.8	31.9	50.0	50.0						
0.4B *ft* (ours)	37.7	53.7	26.5	20.8	59.0	56.6	138.2	76.6	77.4	76.9	38.6	40.0
1.0B *ft* (ours)	38.4	56.6	29.0	22.8	59.7	58.8	142.8	85.1	82.6	78.0	**41.2**	44.7
3.5B *ft* (ours)	39.2	56.9	32.0	26.0	61.3	62.6	141.6	81.8	84.5	77.7	38.2	43.7
11.2B *ft* (ours)	**40.7**	**59.1**	36.8	29.6	62.3	63.9	141.3	83.1	**90.7**	79.3	39.4	46.3
0.4B (ours)	34.1	44.3	34.1	25.9	54.0	53.6	28.2	24.5	12.8	$24.8^{(31.3)}$	$15.5^{(18.9)}$	$0.8^{(1.7)}$
1B (ours)	35.3	47.4	35.4	27.1	54.7	54.1	21.3	17.9	9.4	$17.0^{(34.1)}$	$13.1^{(20.0)}$	$0.8^{(5.6)}$
3.5B (ours)	35.8	48.3	36.7	28.7	53.0	54.3	28.7	24.9	14.8	$24.4^{(38.2)}$	$16.9^{(20.6)}$	$0.8^{(24.1)}$
11.2B (ours)	36.6	50.9	39.5	29.8	52.7	53.5	30.5	26.1	15.6	$24.3^{(39.3)}$	$17.6^{(20.1)}$	$0.9^{(26.6)}$

Table 2: Downstream task performance evaluated on Multi30K Task 1 (translation, mean over the Test2016, Test2017, and AmbiguousCOCO splits), CoMMuTE translation and disambiguation (En→{De, Fr, Es, Ru, Zh}), COCO Karpathy (En captioning), Multi30K Task 2 (captioning), and XM3600 (captioning, Unseen contains {Fr, Es, Ru, Zh}). We report BLEU for translation, accuracy (Acc.) for disambiguation, CIDEr for captioning. **Bold** indicates best results, rows marked with * are evaluated by us, and values with a superscript number in braces are evaluated with a prefix.

P and seen training samples S ($\beta_1 - \beta_2 = -0.18$, 95% CI [-0.48, 0.13]) are not distinctive. While larger models are known to perform well on zero-shot tasks, this is often attributed to potential dataset contamination (Radford et al. 2019). Our findings suggest that generalization is not merely an artifact of pre-training data contamination. Instead, overall model capacity and the quantity of observed, problem-related training data also play a critical role.

During inference, even though the captioning CE loss for unseen languages decreases, the models consistently fail to produce text in the intended target language. Instead, they default to En or De, the two languages encountered during training for captioning. We found that adding a small prefix to the decoder seeds the output of the model. The effect is visualized in Figure 4 and shows the generation of captions without prior exposure to captioning data in those languages. Qualitative examples can be found in the appendix.

We extrapolate the second power law to estimate the CE loss values for a larger, highly multilingual model with P=30B, T=3.0, and a fixed compute budget of S=10M. We predict that this model could achieve a CE loss of 1.92 with a 95% prediction interval (PI) [1.65, 2.23], 1.18 with a 95% PI [0.89, 1.57], and 0.71 with a 95% PI [0.63, 0.80] on UC, ST, and SC, respectively.

Key findings. The insights of this scaling law study can be summarized as follows: CE loss is predicted by initial multilinguality T, model size P, and seen training samples S. For captioning with full coverage (SC), P and S contribute comparably; for translation with only translation supervision (ST), S dominates; and for captioning with only translation supervision (UC), all three matter. Our results indicate that scaling reduces, but does not eliminate, the need for task-language supervision.

Limitations. This study has several limitations. First, our scaling-law analysis is based on only 16 experimental configurations, limiting the predictive capability of our findings. Second, the derived power laws are specific to our experimental setup. Additionally, other factors could influence the parameters, including: the number of languages that the model has to learn, the extensiveness of the pre-training, the synthetic nature of our data and potential style and domain gaps, the difficulty of tasks, and the effect of multiple tasks.

5 Downstream Tasks

To evaluate if the scaling laws transfer, we train on a mix of downstream tasks designed to enhance multilingual translation and captioning.

Training setup. Starting from the models pre-trained for 10K steps, we train for an additional 5K steps using a similar setting. The image resolution is set to 768 px, the batch size to 256, and the learning rate to $5e^{-5}$.

Evaluation setup. We evaluate our models on three tasks: image captioning, multimodal machine translation, and lexical disambiguation. Image captioning is assessed on COCO

Karpathy (Chen et al. 2015; Karpathy and Fei-Fei 2017) (5K images with five En captions), Multi30K (Task 2, 1K images with five En and five De captions) (Elliott et al. 2016), and XM3600 (Thapliyal et al. 2022) (3.6K images, captions in 36 languages) evaluated with CIDEr (Vedantam, Zitnick, and Parikh 2015) using the `pycocoevalcap` toolkit[3]. Following Futeral et al. (2025b), we apply segmentation with `stanza` (Qi et al. 2020) for languages without word boundaries. Multimodal machine translation is assessed with BLEU (Papineni et al. 2002) (via SacreBLEU (Post 2018)) on the Multi30K (Task 1, Test2016, Test2017, AmbiguousCOCO splits) (Elliott et al. 2016, 2017; Barrault et al. 2018) and on CoMMuTE (Futeral et al. 2023, 2025a) (310 translations with images for context). Finally, lexical disambiguation is assessed with accuracy on CoMMuTE.

5.1 Results

We perform OLS regressions in $\log_{10}$ space to model downstream task performance as a function of the CE loss on our UC, ST, and SC test splits. The resulting trend curves are plotted in Figure 5. We observe a strong to moderate fit for most downstream tasks indicating that a lower CE loss on the UC, ST, and SC test splits generally translates to better downstream task performance. For En tasks and De tasks with full task-language coverage, we observe weaker fits. This is likely because performance has begun to plateau, approaching or even exceeding state-of-the-art results on XM3600 and Multi30K De captioning. The captioning task on unseen languages in the Multi30K dataset shows a moderate fit (R^2=0.52). This indicates that the CE loss on the UC task is likely a decent predictor for downstream task performance. However, these interpretations should be taken with caution due to the limited number of measurements.

The downstream task performance is detailed in Table 2. More detailed results can be found in the supplementary material. To put our experimental results into perspective, we compare to a combination of BLIP-2 (Li et al. 2023) and NLLB-3.3B (Costa-jussà et al. 2022) with context-enhanced translation, referred to as Baseline-6B. For captioning, we include PaliGemma-3B (Beyer et al. 2024) and Florence-2-L (Xiao et al. 2024). For translation, we use NLLB-3.3B, its multimodal extension ZeroMMT-3.3B (Futeral et al. 2025a), Multilingual Open Flamingo (MOF) (Futeral et al. 2025b), and multilingual VGAMT (Futeral et al. 2023). Furthermore, we include three state-of-the-art multilingual VLMs as references: Pixtral-12B (Agrawal et al. 2024) and Gemma-3-12B (Gemma Team 2025). Note that baseline models vary in their degree of exposure to the downstream training data.

Before fine-tuning, the performance across all benchmarks of our models is relatively weak but improving with scale. Translation performance is slightly worse than our baseline. However, the captioning metrics appear artificially low, likely due to a style and domain mismatch. A language-specific prefix (see Figure 4) resolves the complete failure on unseen-language captioning tasks (0.9 CIDEr without and 26.6 CIDEr with prefix for the 11.2B model on XM3600 unseen) while also boosting CIDEr scores for En and De. The "step down" in captioning performance between 1.0B and 3.5B, along with the slight drop in performance of 1.0B in translation benchmarks, suggests that learning a multilingual embedding layer is more difficult than learning multimodal alignment. The gap does not transfer to fine-tuned models, however.

Unsurprisingly, fine-tuning on the combined downstream task dataset leads to substantial scalable performance improvements across nearly all benchmarks. Our model achieves the best performance for Multi30K Translation (40.7 and 59.1 BLEU for En→{De, Fr}, respectively), and image captioning in De on Multi30K (90.7 CIDEr) and XM3600 (41.2 CIDEr), outperforming both specialized and more-capable foundation models. On unseen languages in XM3600, our models are competitive, even surpassing Baseline-6B, and perform only slightly worse than Gemma-3-12B, despite not relying on extensive multilingual multimodal pre-training. The overall best performance is achieved by Pixtral-12B (50.5 CIDEr).

The BLEU scores for CoMMuTE translation slightly decrease after fine-tuning. We attribute this to the dataset containing translations that stylistically differ. Gemma-3-12B outperforms all other models on this task including the NLLB translation model. Similarly, the Gemma-3 models are excellent at resolving lexical ambiguities on CoMMuTE followed by Pixtral-12B. Our models fall behind, as they are trained on synthetic data and only a small dataset with real annotations. The disambiguation accuracy stays constant for pre-training, while it increases with scale for fine-tuning. MOF achieves 66.5% mean accuracy from pre-training on web-crawled data but shows reduced Multi30K translation performance. Our approach features 63.9% mean accuracy while maintaining good results for Multi30K translation. Overall, our findings highlight the need for small high-quality datasets with full task coverage in each language addressing these ambiguities, while the pre-training dataset can be of lower quality or even incomplete.

6 Conclusion

We presented scaling laws for generalization from multimodal machine translation to multilingual image captioning, demonstrating how transfer performance scales with the multilinguality of the base model, the model size, and the amount of training data. While captioning in languages encountered only in the translation task still requires a language prefix in a zero-shot setting, our results highlight that these factors strongly influence how well encoder-decoder VLMs extend learned language capabilities to unseen task-language combinations. Fine-tuning removes the need for explicit prefixes and yields competitive performance across downstream tasks. Our insights can help practitioners create multilingual datasets more efficiently and make informed trade-offs between model size, multilingual pre-training, number of training samples, and task coverage. Future work can investigate the interactions of more than two tasks and the extension of our findings to decoder-only VLMs, potentially leading to better, more versatile multilingual models.

[3]https://github.com/salaniz/pycocoevalcap

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany under grant no. 01IS23004B RIDMI and 01IS22094C WEST-AI. Computational resources were provided by the German AI Service Center WestAI.

References

Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D.; Chudnovsky, J.; Costa, D.; De Monicault, B.; Garg, S.; Gervet, T.; et al. 2024. Pixtral 12B. *CoRR*, abs/2410.07073.

Alayrac, J.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; et al. 2022. Flamingo: A visual language model for few-shot learning. In *NeurIPS*.

Barrault, L.; Bougares, F.; Specia, L.; Lala, C.; Elliott, D.; and Frank, S. 2018. Findings of the third shared task on multimodal machine translation. In *3rd Conference on Machine Translation (WMT)*, 304–323.

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; et al. 2024. PaliGemma: A versatile 3B VLM for transfer. *CoRR*, abs/2407.07726.

Carlsson, F.; Eisen, P.; Rekathati, F.; and Sahlgren, M. 2022. Cross-lingual and multilingual CLIP. In *LREC*, 6848–6854.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 3558–3568.

Chen, G.; Hou, L.; Chen, Y.; Dai, W.; Shang, L.; Jiang, X.; Liu, Q.; Pan, J.; and Wang, W. 2023. mCLIP: Multilingual CLIP via cross-lingual transfer. In *ACL*, 13028–13043.

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2818–2829.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, 8440–8451.

Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; et al. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

de Souza, L. R.; Almeida, T. S.; Lotufo, R. A.; and Nogueira, R. F. 2024. Measuring cross-lingual transfer in bytes. In *NAACL*, 7526–7537.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library. *CoRR*, abs/2401.08281.

Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; and Specia, L. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *2nd Conference on Machine Translation (WMT)*, 215–233.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30K: Multilingual English-German image descriptions. In *VL@ACL*.

Fernandes, P.; Ghorbani, B.; Garcia, X.; Freitag, M.; and Firat, O. 2023. Scaling laws for multilingual neural machine translation. In *ICML*, 10053–10071.

Firat, O.; Sankaran, B.; Al-Onaizan, Y.; Yarman-Vural, F. T.; and Cho, K. 2016. Zero-resource translation with multilingual neural machine translation. In *EMNLP*, 268–277.

Fodor, J. A.; and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2): 3–71.

Futeral, M.; Schmid, C.; Laptev, I.; Sagot, B.; and Bawden, R. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *ACL*, 5394–5413.

Futeral, M.; Schmid, C.; Sagot, B.; and Bawden, R. 2025a. Towards zero-shot multimodal machine translation. In *NAACL*, 761–778.

Futeral, M.; Zebaze, A. R.; Suarez, P. O.; Abadji, J.; Lacroix, R.; Schmid, C.; Bawden, R.; and Sagot, B. 2025b. mOSCAR: A large-scale multilingual and multimodal document-level corpus. In *ACL*, 3461–3494.

Gee, L.; Zugarini, A.; Rigutini, L.; and Torroni, P. 2022. Fast vocabulary transfer for language model compression. In *EMNLP*, 409–416.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Gemma Team. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Ghorbani, B.; Firat, O.; Freitag, M.; Bapna, A.; Krikun, M.; Garcia, X.; Chelba, C.; and Cherry, C. 2022. Scaling laws for neural machine translation. In *ICLR*.

Gogoulou, E.; Ekgren, A.; Isbister, T.; and Sahlgren, M. 2022. Cross-lingual transfer of monolingual models. In *LREC*, 948–955.

Hernandez, D.; Kaplan, J.; Henighan, T.; and McCandlish, S. 2021. Scaling laws for transfer. *CoRR*, abs/2102.01293.

Hirasawa, T.; Bugliarello, E.; Elliott, D.; and Komachi, M. 2023. Visual prediction improves zero-shot cross-modal machine translation. In *8th Conference on Machine Translation (WMT)*, 522–535.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; et al. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F. B.; Wattenberg, M.; Corrado, G.; Hughes, M.; and Dean, J. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 339–351.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Karpathy, A.; and Fei-Fei, L. 2017. Deep visual-semantic alignments for generating image descriptions. *TPAMI*, 39(4): 664–676.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 3337–3345.

Lake, B. M.; and Baroni, M. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985): 115–121.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742.

Libovický, J.; Rosa, R.; and Fraser, A. 2019. How language-neutral is multilingual BERT? *CoRR*, abs/1911.03310.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *NeurIPS*.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual denoising pre-training for neural machine translation. *TACL*, 8: 726–742.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *ICLR*.

Mitzalis, F.; Caglayan, O.; Madhyastha, P.; and Specia, L. 2021. BERTGen: Multi-task generation through BERT. In *ACL/IJCNLP*, 6440–6455.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; et al. 2023. Crosslingual generalization through multitask finetuning. In *ACL*, 15991–16111.

Onoe, Y.; Rane, S.; Berger, Z.; Bitton, Y.; Cho, J.; Garg, R.; Ku, A.; Parekh, Z.; Pont-Tuset, J.; Tanzer, G.; Wang, S.; and Baldridge, J. 2024. DOCCI: Descriptions of connected and contrasting images. In *ECCV*, volume 15118, 291–309.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 311–318.

Pires, T.; Schlinger, E.; and Garrette, D. 2019. How multilingual is multilingual BERT? In *ACL*, 4996–5001.

Post, M. 2018. A call for clarity in reporting BLEU scores. In *3rd Conference on Machine Translation (WMT)*, 186–191.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *ACL*, 101–108.

Qiu, C.; Oneata, D.; Bugliarello, E.; Frank, S.; and Elliott, D. 2022. Multilingual multimodal learning with machine translated text. In *EMNLP*, 4178–4193.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Schwenk, H.; Wenzek, G.; Edunov, S.; Grave, E.; Joulin, A.; and Fan, A. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *ACL/IJCNLP*, 6490–6500.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 843–852.

Thapliyal, A. V.; Pont-Tuset, J.; Chen, X.; and Soricut, R. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *EMNLP*, 715–729.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, 38–45.

Wu, H.; and Wang, H. 2007. Pivot language approach for phrase-based statistical machine translation. *Mach. Transl.*, 21(3): 165–181.

Wu, S.; and Dredze, M. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *EMNLP-IJCNLP*, 833–844.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; et al. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 4818–4829.

Zheng, W.; Pan, W.; Xu, X.; Qin, L.; Yue, L.; and Zhou, M. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *EMNLP*, 7725–7738.