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

Julian Spravil, Sebastian Houben, Sven Behnke

Project Page

spravil.com/projects/caption_via_translation/

TL;DR: Scaling model size, training samples, and multilinguality enables **image captioning in unseen languages via translation as auxiliary task**, however, fine-tuning with full task-language coverage remains essential.

1. Introduction & Motivation

- Current advances in multilingual multimodal modeling rely either on **expensively pretrained multilingual language models** or **large-scale datasets**
- Real-world datasets are **incomplete by nature:** Not all tasks are covered in all languages!
- **The dynamics of multilingual cross-task generalization, particularly at scale, remain unexplored**

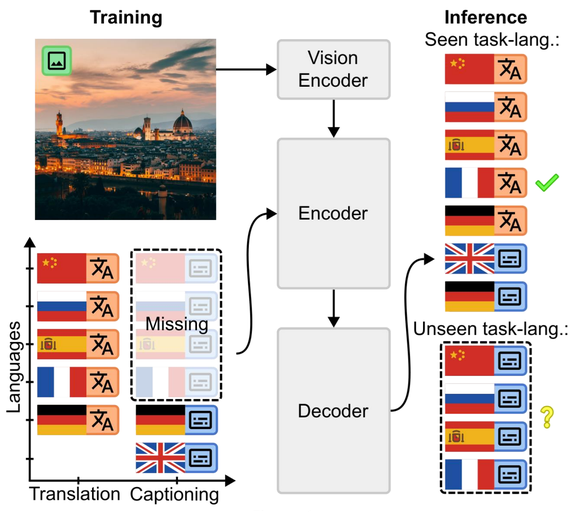

Figure 1

Goal: Investigate **systematic generalization** [0] within a **realistic multimodal setting** using **partial pre-trained encoder-decoder transformers** and a **standard training method**

2. Method

Intentionally Incomplete Synthetic Dataset

- Data source: **CC12M images** and **CCMatrix translation pairs**
- Models used for dataset creation: Vision-language (LLaVA-NeXT [1]), machine translation (NLLB-3.3B [2]), and contrastive (CLIP-ViT/B-16 [3]) models
- Covered languages: **English** (En), **German** (De), **French** (Fr), **Spanish** (Es), **Russian** (Ru), and **Chinese** (Zh)
- 10M Images, 32M captions (En, De), 105M translations (En→{De, Fr, Es, Ru, Zh})
- **Eval dataset:** 4.4K images with **captioning in Es and Zh (UC)**, **translation from En to Es and Zh (ST)**, and **captioning in En and De (SC)**

Models

- **Encoder-Decoder Model:** Based on **Florence-2** [4] and **Gemma-2** [5]
- Florence-2 base (0.4B), Florence-2 large (1.0B) with **Gemma-2 tokenizer**
- Gemma-2 (3.5B and 11.2B) with Florence-2 large encoder and inserted cross-attention layers

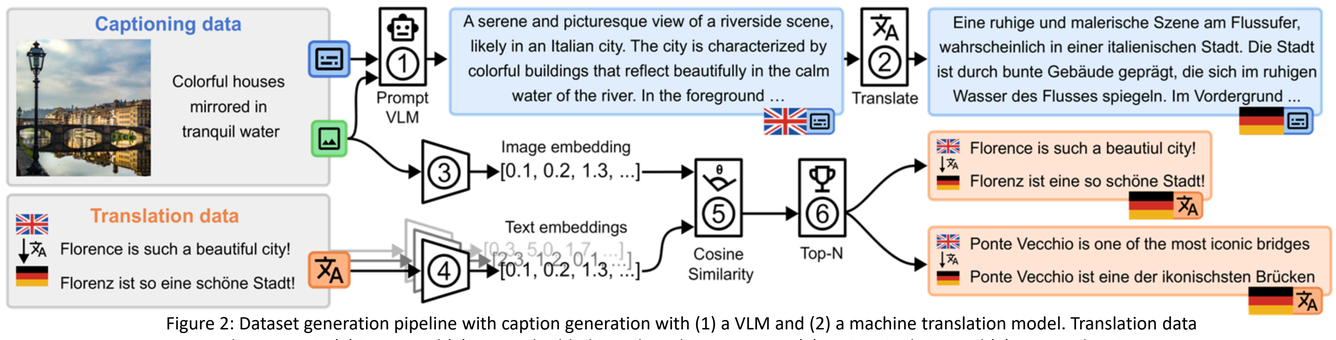

Figure 2: Dataset generation pipeline with caption generation with (1) a VLM and (2) a machine translation model. Translation data alignment via (3) image and (4) text embedded in a shared vector space, (5) cosine similarity and (6) top-N selection.

3. Multilingual Cross-Task Generalization at Scale

- **Fitted power law:** CE loss is predicted by **initial multilinguality T**, **model size P**, and **seen training samples S**

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

Key Insights (compare with the Figure on the right)

- **UC:** For captioning with only translation supervision, all three variables matter
- **ST:** For translation with only translation supervision (ST), S dominates
- **SC:** For captioning with full coverage (SC), P and S contribute comparably

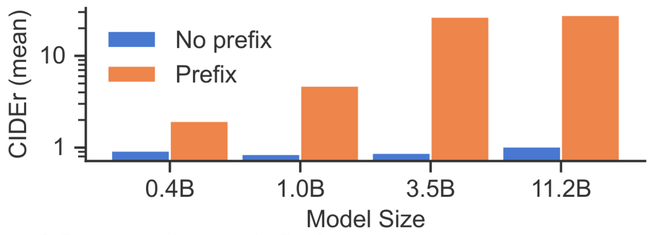

Figure 3: Effect of adding a prefix (Fr: “La photo montre”, etc.) to the decoder input to unlock zero-shot captioning on XM3600.

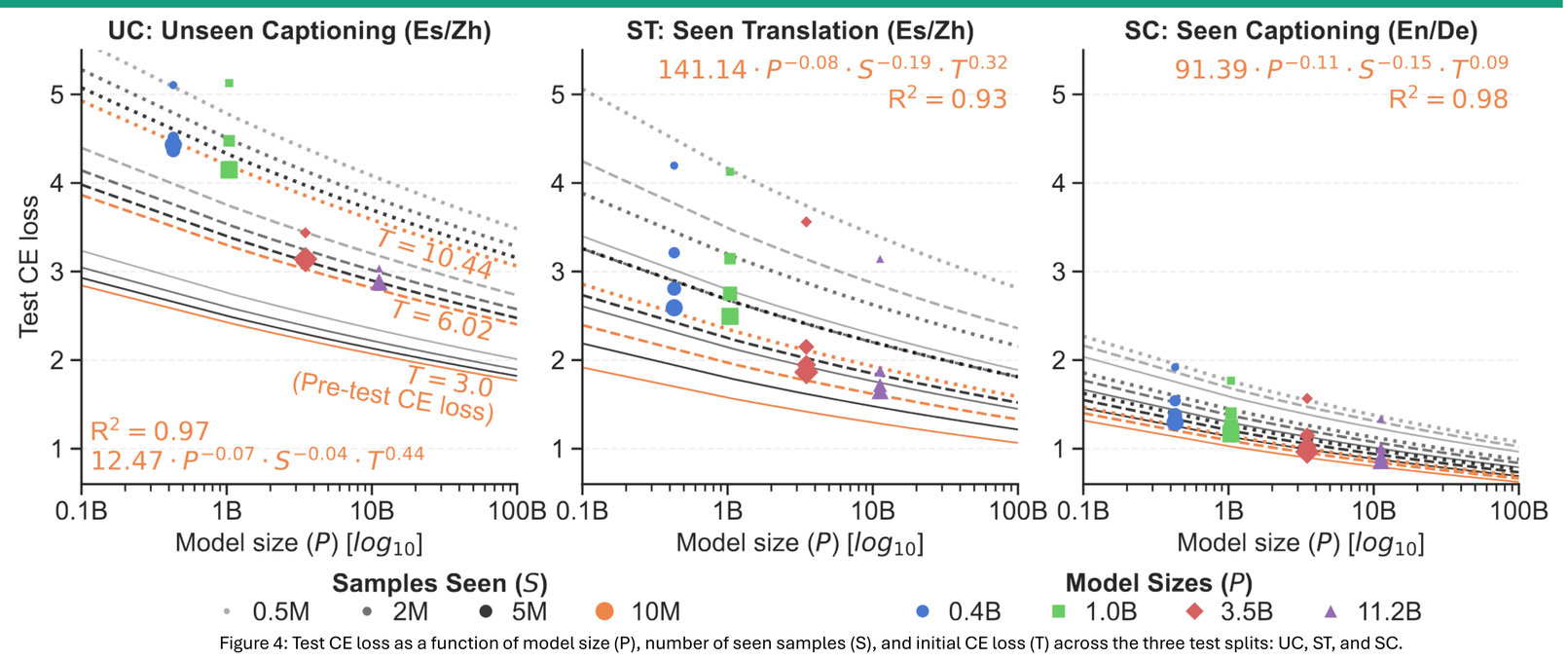

Figure 4: Test CE loss as a function of model size (P), number of seen samples (S), and initial CE loss (T) across the three test splits: UC, ST, and SC.

4. Transfer to Downstream Tasks

Finetuning Dataset

- The downstream dataset consists of **Multi30k (task 1 translation, task 2 captioning), Image Paragraphs, DOCCI, and COCO Karpathy**
- **Missing languages are added via machine translation** using NLLB-3.3B
- In total, the dataset contains 166K images with 1.6M captions and 145K translations with full task-lang coverage

Some Insights

- **SOTA German image captioning on Multi30K** and **competitive performance** across the task selection
- **CE loss is a useful proxy for downstream performance:** lower CE loss generally predicts better downstream results, even for unseen languages for captioning
- **Fine-tuning is necessary** to unlock all unseen task-language pairs at the studied scales

Takeaway: Building a multilingual model? Include proxy tasks for the target languages for pre-training and build a small fine-tuning dataset with full task-language coverage!

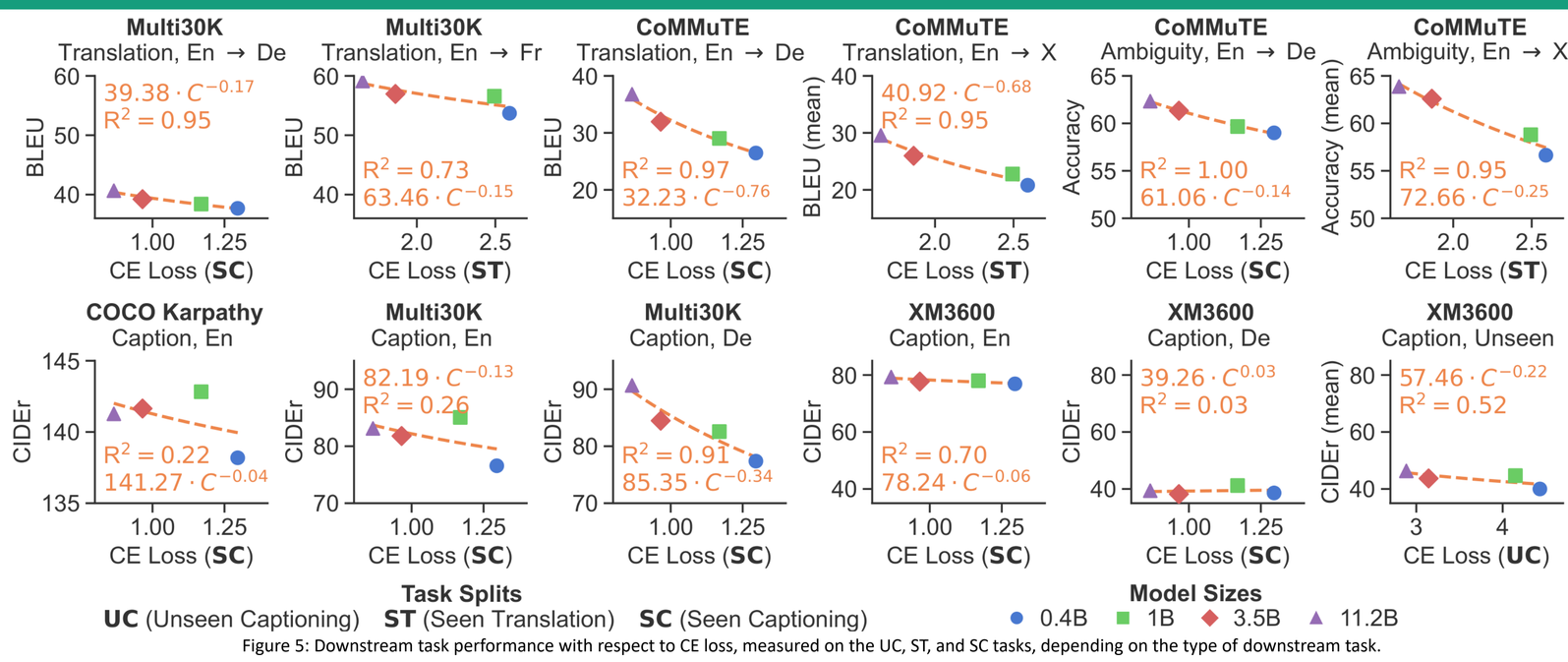

Figure 5: Downstream task performance with respect to CE loss, measured on the UC, ST, and SC tasks, depending on the type of downstream task.

5. Examples (11.2B Model)

EN: The image shows a night view of the gardens by the bay in singapore, featuring the iconic supertree grove.
FR: La photo montre une vue de nuit de l'hôtel Marina Bay Sands à Singapour, qui est célèbre pour son toit en forme de pétale de lotus.
FT-COCO-ES: Los árboles están iluminados con luces púrpuras. (The trees are illuminated with purple lights.)

ES: La imagen muestra una cascada de agua que cae desde el techo de un edificio, creando un efecto de lluvia artificial.
ZH: 图为 新加坡滨海湾花园的空中花园，该花园以其独特的玻璃穹顶结构而闻名。
FT-COCO-ZH: 喷泉从天花板上喷出。 (The fountain spouted from the ceiling.)

EN: The iconic marina bay sands hotel in singapore, a marvel of modern architecture.
DE: Das ikonische Marina Bay Sands Hotel in Singapur, ein berühmtes Wahrzeichen, das für seine unverwechselbare Architektur bekannt ist.
FT-COCO-ZH: 两座摩天大楼在阳光明的一天。 (Two skyscrapers on a sunny day.)

FT-COCO-EN: Three pelicans standing on a dock next to the water.
FT-COCO-DE: Drei Pelikane stehen auf einem Steg. (Three pelicans are standing on a pier.)
FT-COCO-FR: Trois pélicans debout sur un quai. (…)
FT-COCO-RU: Три пеликана стоят на пирсе. (Three pelicans are standing on the pier.)

PREFIX inserted for condition emergence
FT-COCO refers to the fine-tuned model with prompted COCO caption format

6. Conclusion

- **Transfer to multilingual image captioning improves predictably as you increase base model multilinguality, model size, and training data**
- In zero-shot captioning for languages only encountered in translation data, an explicit **language prefix is needed**
- **Fine-tuning removes the prefix requirement** and achieves competitive downstream performance
- The results guide efficient multilingual dataset construction and trade-offs among model capacity, multilingual pre-training breadth, sample counts, and task coverage
- **Future work:** Multi-task interactions beyond two tasks and testing whether these findings extend to decoder-only VLMs

References
[0] Fodor, J. A.; and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1-2): 3–71.
[1] Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In NeurIPS.
[2] Costa-jussà, M. R.; Cross, J.; C¸ elebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; et al. 2022. No language left behind: Scaling human-centered machine translation. CoRR, abs/2207.04672.
[3] 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.
[4] 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.
[5] Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. CoRR, abs/2408.00118.

Fraunhofer IAIS
UNIVERSITÄT BONN
Hochschule Bonn-Rhein-Sieg University of Applied Sciences
LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SPONSORED BY THE
Federal Ministry of Research, Technology and Space