Do captions in different languages produce different images?: Efficiently training Multilingual Diffusion

Aadit Deshpande, Liam Walker, Sreenidhi Ganachari

Carnegie Mellon University

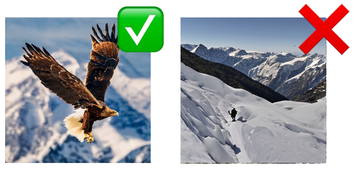
Dataset and Code:
https://github.com/Aadit3003/genai-multilingual-tti

Text-to-Image Challenges

- Existing T2I diffusion models are often limited to English, restricting accessibility and global adoption.
- Our implementation of AltDiffusion integrates a multilingual text encoder and fine-tuned U-net into a pre trained English Stable diffusion model to support multiple languages.

EN FR

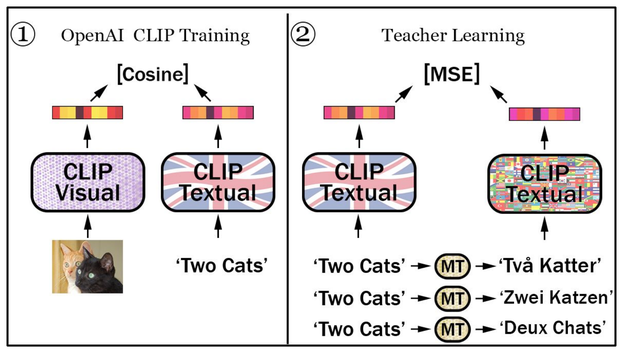
Colorful rainbow smoothie **chia bowl**



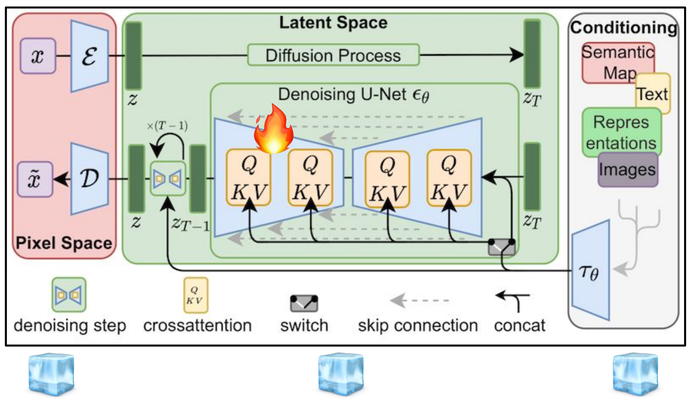
EN DE

Golden eagle **soaring** over snowy peaks

Multilingual Training



Stage1: **Teacher Learning**
(Text Encoder)



Stage2: **Concept Alignment**
(U-Net with LoRA)

Stage 1: Teacher Learning

Enhance language capability of CLIPText Encoder

- CLIP ViT-H/14 model
- Teacher-Learning (25k FR/DE sentences with parallel EN transl.)
- Minimize MSE Loss

Stage 2: Concept Alignment

Align U-Net with new text encoder embeddings

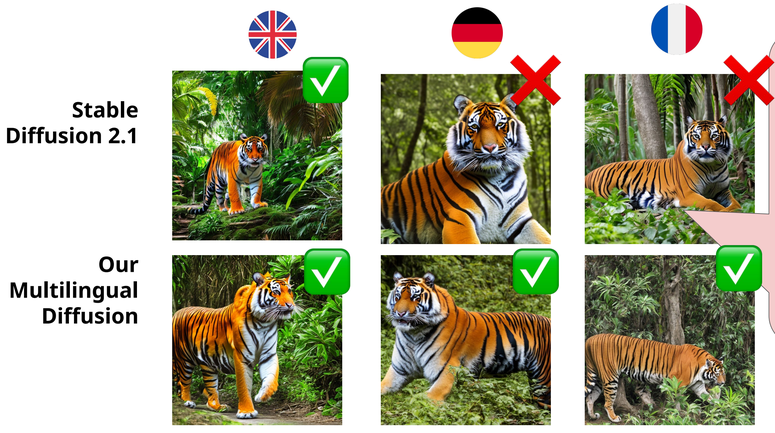
- Freeze the Text Encoder, VAE
- **Add Adapters to U-Net**
- **Inject LoRA trainable matrices: in the U-Net: K, Q, V, Out_Proj**
- Fine-tune with **LoRA rank 4**
- 15000 train steps on Train Set (6k) with EN/DE/FR

Qualitative Analysis

Prompt - Darth Vader in a blue christmas sweater sipping coffee near the fireplace

Complex Subordinate Clauses: Subordinate clauses (**blue christmas sweater, near the fireplace**) are hard for SD2.1 to combine, especially in French

Prompt - A majestic Bengal tiger with vibrant orange fur, **stalking** through a lush tropical rainforest dappled with sunlight



Information Loss: For French and German, SD2.1 Captures a generic "Tiger", but **ignores aspects like "stalking"**

Experiments and Results

Evaluation Dataset Creation

- Subset of WIT Dataset
  - *Wikipedia-based Image Text*
- **3 languages: English (EN), French (FR), German (DE)**
- Filtering Steps for high quality
  - Word Count, Image Size
  - CLIP Scores >= 20
- Train/Test Splits: 6k vs 1.5k

Evaluate Diffusion Models

- **Cascading Monolingual**
  - NLLB-200 (translate) then
  - Stable Diffusion 2.1 (generate)
- **End-to-end FT Multilingual**
- 512x512 inference resolution
- Classifier-free Guidance Scale 7.5 (50 DDIM steps)

NLLB-200 + Stable Diffusion v2.1

Language	FID(↓)	IS(↑)
EN	**1.08**	11.33
DE	1.16	**11.45**
FR	1.30	11.17

Our Multilingual Diffusion

Language	FID(↓)	IS(↑)
EN	0.99	11.72
DE	1.04	**12.42**
FR	**0.95**	11.63

Conclusion

- Our multilingual end-to-end model **outperforms vanilla** Stable Diffusion in FID and IS score across **FR and DE,** while **still improving on EN**.
- Our method better captures nuances likely to be "lost in translation"

References

- Ye, Fulong, et al. "**Altdiffusion**: A multilingual text-to-image diffusion model." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 7. 2024.
- Carlsson, Fredrik, et al. "**Cross-lingual and multilingual clip**." *Proceedings of the thirteenth language resources and evaluation conference*. 2022.