MultiFusion: Fusing Pre-Trained Models for Multi-Lingual, Multi-Modal Image Generation

Marco Bellagente[4*+], Manuel Brack[2,3*], Hannah Teufel[1*], Felix Friedrich[3,6], Björn Deiseroth[1,3,6], Constantin Eichenberg[1], Andrew Dai[1], Robert J.N. Baldock[1], Souradeep Nanda[5*], Koen Oostermeijer[1], Andres Felipe Cruz-Salinas[1], Patrick Schramowski[2,3,6,8], Kristian Kersting[2,3,6,7#], Samuel Weinbach[1#]

marco.bellagente@gmail.com, brack@cs.tu-darmstadt.de, hannah.teufel@aleph-alpha.com

[1]Aleph Alpha, [2]German Research Center for Artificial Intelligence (DFKI), [3]Computer Science Department, TU Darmstadt, [4]Stability AI, [5]University of Texas, [6]Hessian AI, [7]Center for Cognitive Science, TU Darmstadt, [8]LAION

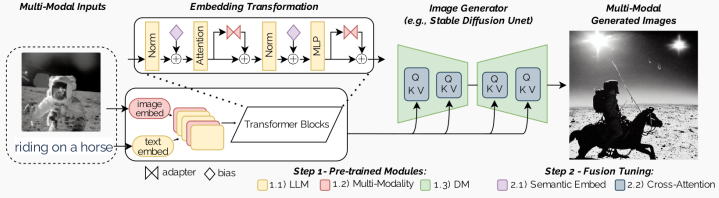
+Work done while at Aleph Alpha, *Equal Contribution, #Equal supervision

I. Method

How to enable the **intuitive, multilingual expression** of **complex** and **nuanced concepts** in **Diffusion Models** both **compute efficiently** and **without multilingual downstream** training data?

Interleaved multimodal, multilingual prompting

Fusion of pretrained models to a modular encoder, which transfers capabilities to the downstream diffusion model



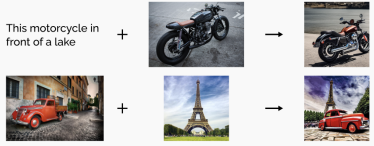
To enable multimodal, multilingual prompting we replace Stable Diffusions Clip encoder by a custom modular one.

1.1) An autoregressive **LLM pre-trained** on 5 languages enables **multilingualism**.
1.2) Extending the LLM by an **image prefix** as well as adapters enables **multimodality**.
2.1) **Finetuning** the **biases** of the LLM provides embeddings, which capture the **semantic meaning** of the text prompt, thus simplifying the learning of mapping from embeddings to image outputs.
2.2) To **align** the **pre-trained Stable Diffusion** model (1.4) with the embeddings of our modular encoder, we retrain the conditioning by finetuning the **cross-attention** weights.

II. Application

Image Composition:

MultiFusion increases **expressiveness** in **composition** through arbitrary and flexible prompting of image and text sequences.



Negative Prompting:

Negative prompting with images enables a more **powerful suppression** than through text.

Style Modification:

MultiFusion enables **simple style transfer** through one **reference** image capturing all the facets of a unique style such as as color palette, composition, contrast, etc., making elaporate prompts obsolete. Additionally, MultiFusion enables **highly individual** prompting such as "In the style of a picture I drew".

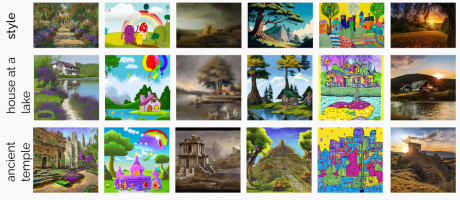
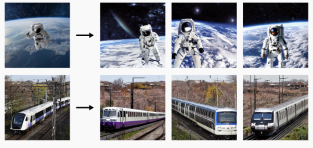


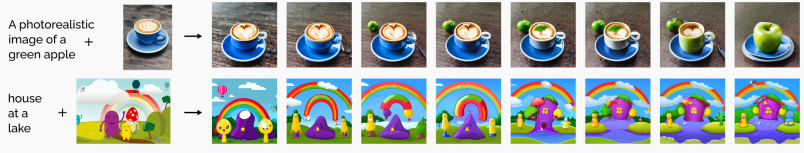
Image Variation:

MultiFusion produces **meaningful image variations** without the need for inversion or renoising of the input image



Attention Manipulation:

Attention Manipulation allows us to **weight image and text** tokens at inference time and thus guide their **influence** on the resulting generation.



III. Evaluation

Image fidelity and image-text alignment:

We measure **image fidelity** and **image text-alignment** using the standard metrics FID-30K and Clip Scores. We find that MultiFusion prompted with text only performs **on par with stable diffusion** despite extension of the Encoder to support multiple languages and modalities.

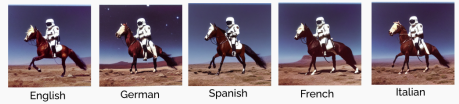
	FID-30K ↓			CLIP Score (Text-to-Image) ↑	
Guidance Scale	SD v1.5	MF (Text)	MF (Multimodal)	MF (Image)	SD v1.5	MF
8.0	14.62	14.19	13.50	8.02	0.31	0.30
6.0	12.73	12.35	10.29	7.18	0.31	0.29
4.0	10.19	9.90	8.53	6.03	0.31	0.29
2.0	9.74	12.21	8.61	6.05	0.30	0.28
1.0	26.09	32.81	24.22	18.93	0.27	0.25

Compositional Robustness through Multimodality:

Image composition is a known **limitation of Diffusion Models**. Evaluating on our new **benchmark MCC-250** we show that interleaved **multimodal prompting** leads to more **compositional robustness** as judged by humans.

Model	Two objects w/ correct colors (%)
Stable Diffusion	29.92
Composable Diffusion	25.59
MultiFusion (text)	21.66
MultiFusion (multimodal)	**58.35**

Multilinguality:



Evaluating the **alignment** of prompt embeddings as well as generated images across **multiple languages** we show that **good embedding alignment** enables the **transfer** of multilingualism to downstream tasks even for task-specific monolingual training data.