universität freiburg

https://github.com/shengwenyuan/mensa_t2i.git

SoTA Text to Image Adapting and Finetuning MensaFood

Andi Alidema, Wenyuan Sheng, Rona Latifaj

Karim Farid and Leonhard Sommer

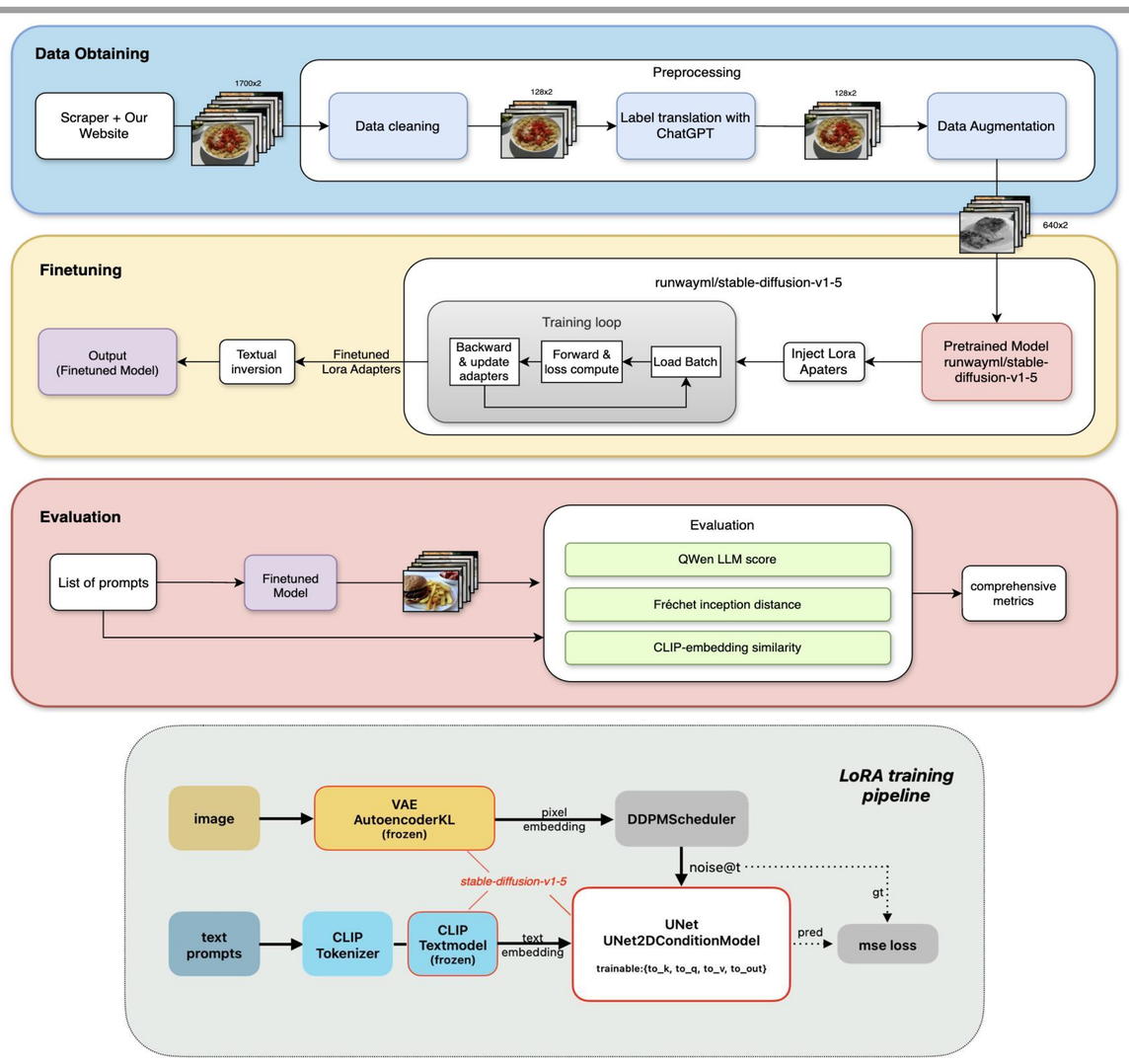
University of Freiburg

Introduction

Let's say you just walked into your first class of the day and are wondering what to eat after. You open the SWFR page and read what the mensa is offering today. Would it not be nice to see what the food would actually look like, prepared the same way it's served in our university mensa specifically? This is the challenge that we are tackling in this project.

The world of AI-generated images has seen massive breakthroughs in the last couple of years. Yet out-of-the-box models do not yet produce images we would call "exactly what the mensa served today." In this "SoTA Mensa Text to Image" project we are teaching these impressive visual generative models to capture specialized domains—like the familiar yet varied dishes served in University Mensa—and provide us with the exact foods we will be served later in the day.

Method



Quantitative Results

Table 1: FID eval

model	FID score
diffusion pretrained	398
lora-v1	**165**
lora-v2	185
lora-v3	234
lora-v4	176

Table 2: CLIP & LLM metrics

images	CLIP score		QWen score (0-100)	
	arith	geo	arith*	geo†
gt dataset	0.3098	0.3084	69.30	54.75
lora-v1 (raw image-text pairs)	**0.3083**	**0.3078**	73.17	68.42
lora-v2 (unique image-text pairs)	0.2858	0.2846	58.00	32.07
lora-v3 (English-text)	0.2805	0.2785	67.38	39.42
lora-v4 (v2 & v3 & image-aug & keyword prefix)	0.3009	0.3001	**78.79**	**78.29**

The evaluation dataset is composed with 10 real mensafood images and corresponding descriptions in DE and EN.

* arithmetic mean of each food object score across eval images

† geometric mean of each food object score across eval images

Table 3: LLM prompt of one food object

Describe	You are my assistant to identify any objects and their {texture} in the image.
Predict	Evaluate if there are '{sliced pepper cucumber salad}' in the image according the criteria: A: there are {pepper cucumber salad}, and {texture} is good. {sliced} are appropriate for the contents. B: there are {pepper cucumber salad}, but {texture} is bad. C: there is {salad}, but not all {pepper cucumber salad} appear. D: no {salad} in the image. Provide a score (0-100) and explanation in JSON format

Limitations

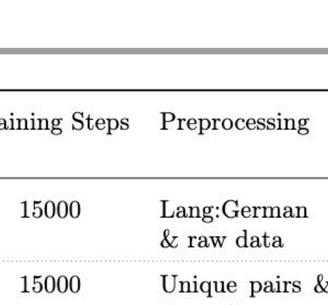
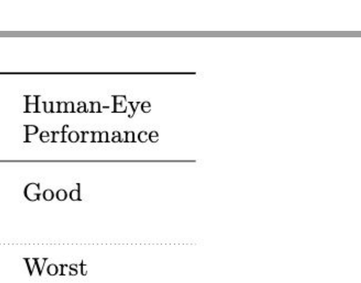
- Compositional accuracy: Model struggles with complex multi-item food arrangements, meat doneness texture, and realistic portion scaling;
- Results may not generalize equally to German prompts.
- Future work could explore: higher-resolution training, adaptive LoRA rank selection to balance expressivity and efficiency, simultaneous fine-tuning of the text encoder and VAE for richer semantic alignment.

Experiments

Version	Resolution	Training Steps	Preprocessing	Human-Eye Performance
v1	256×256	15000	Lang:German & raw data	Good
v2	512×512	15000	Unique pairs & Lang:German	Worst
v3	512×512	30000	Unique pairs & Lang:English	Second worst
v4	512×512	30000	Lang:English & Data Augmentation & UP	Best

Table 1: Comparison of Model Versions

Qualitative Results

Prompt	Model Generated Image	Real Image
Mensafood Breaded pork schnitzel or vegetable schnitzel, roast gravy, and French fries		
Mensafood Pasta-Kreationen aus unserer Pasta-Manufaktur mit verschiedenen Saucen und Toppings		
Mensafood Stuffed peppers, baked potato and sweet-potato wedges, Spanish beans, and side salad		
Mensafood Hamburger TS with beef, cheese, tomato, lettuce, and French fries		
Mensafood Tortellini gefüllt mit Ricotta und SpinatBasilikum-Käsesauce Frühlingszwiebel Tomatenwürfel und geriebener Emmentaler		

References

[1] Cuenca & Paul, Hugging Face Blog, Jan 2023. "Using LoRA for Efficient Stable Diffusion Fine-Tuning."

[2] Heusel *et al.*, *NeurIPS* 2017. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium."

[3] Huang et al., IEEE TPAMI 2025. "T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation."