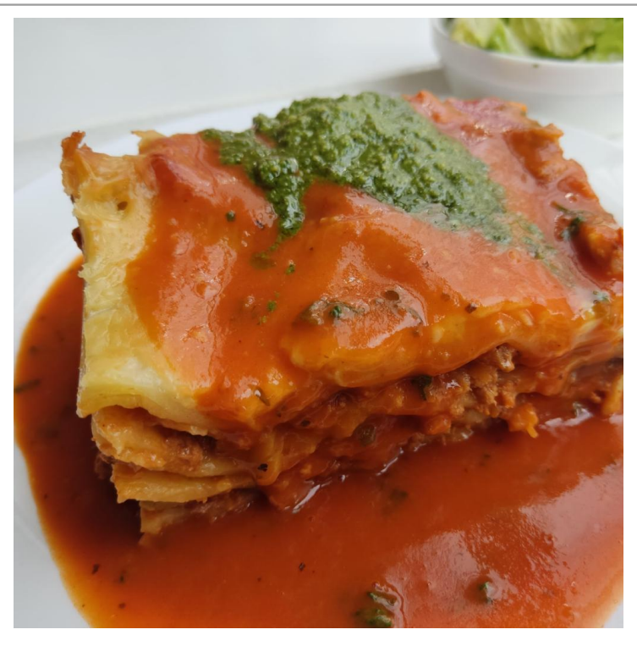
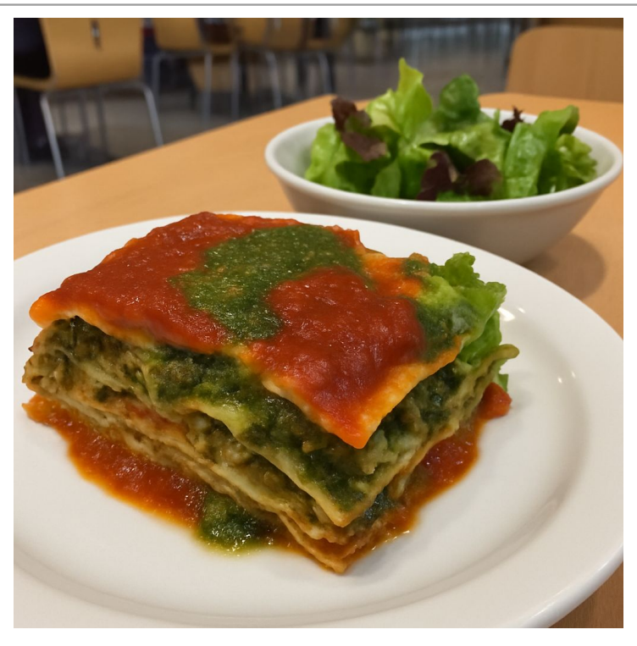
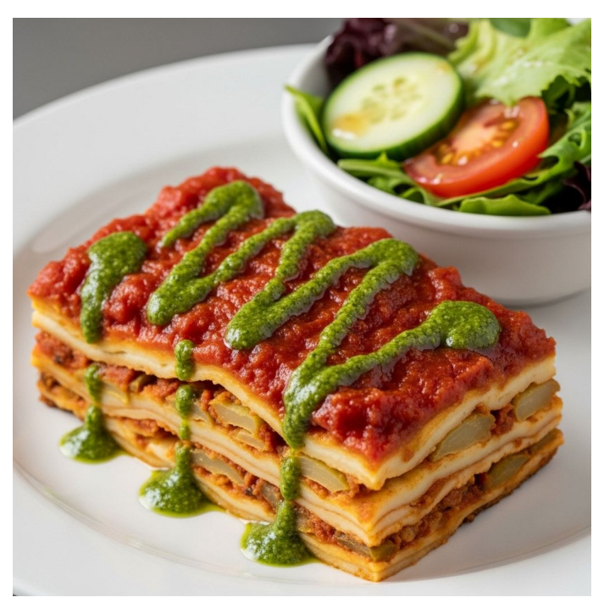
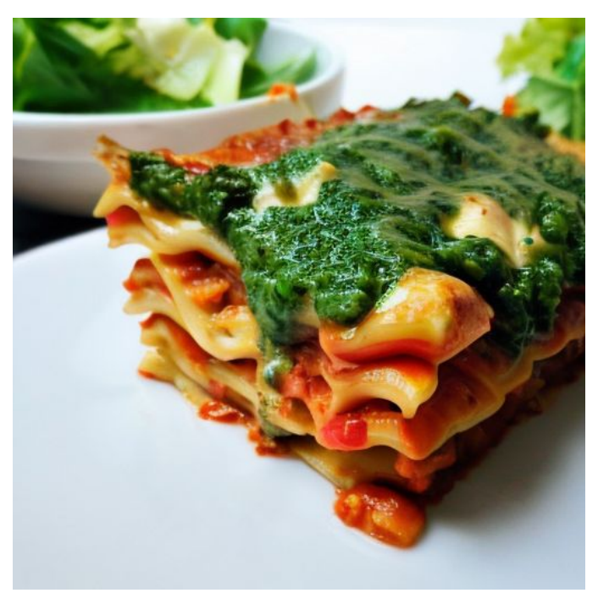
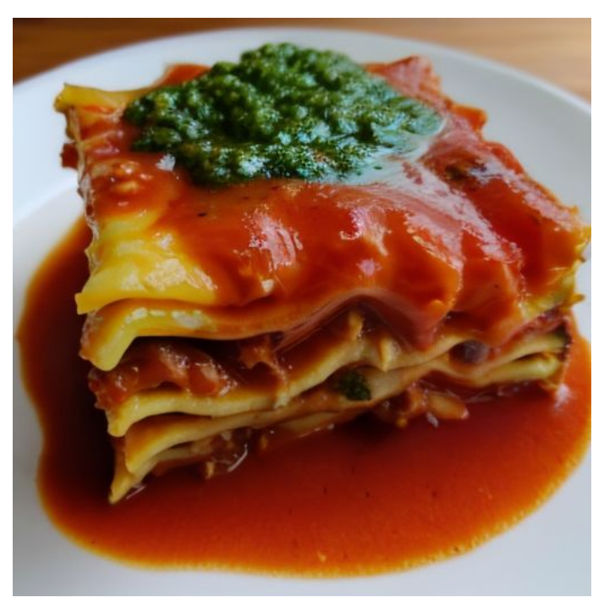
SoTA T2I Adapting and Finetuning - mensa-food

Murat Han Aydoğan - Frederik Lars Melbye - Ali Safarli
University of Freiburg

Karim Farid

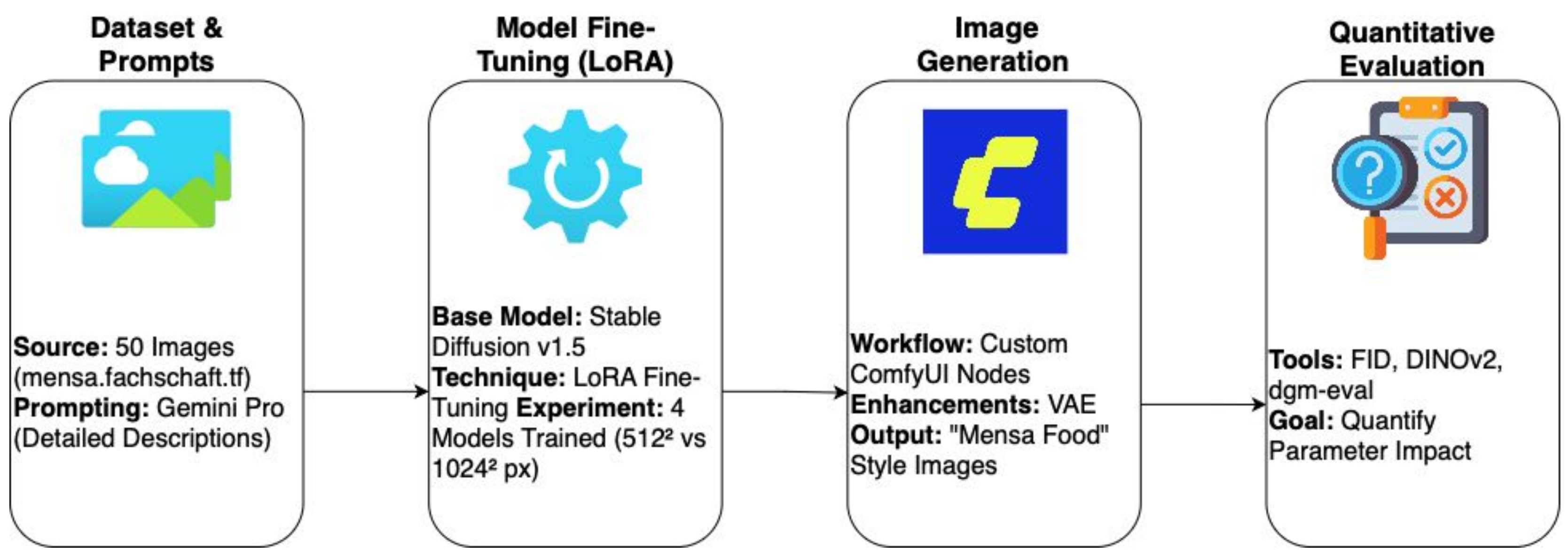
Introduction

While state-of-the-art text-to-image models excel at generating idealized images, their performance in highly specific, real-world domains remains a key challenge. This project investigates adapting Stable Diffusion v1.5 [1] to the niche visual style of German university cafeteria (Mensa) food. Our primary obstacle was the nature of the dataset itself; sourced from authentic photos, it was characterized by inconsistent lighting, poor shot angles, and non-professional composition.
Using a curated set of just 50 images, we employed Low-Rank Adaptation (LoRA) [2] to fine-tune the model, exploring the impact of 512x512 vs. 1024x1024 resolutions [3]. Our results were evaluated qualitatively in a ComfyUI workflow and quantitatively with FD, DINOv2, and dgm-eval [4] metrics. The findings definitively show that LoRA is a highly effective strategy for domain adaptation, proving that a model can successfully learn a specific and imperfect visual style from as few as 50 images to generate coherent and stylistically accurate results.

Real Image	ChatGPT	Gemini	mensa-food-v3	mensa-food-v4
				

Method

Our project followed a systematic pipeline from data collection to quantitative evaluation.



Dataset & Prompts
Source: 50 Images (mensa.fachschaft.tf)
Prompting: Gemini Pro (Detailed Descriptions)

Model Fine-Tuning (LoRA)
Base Model: Stable Diffusion v1.5
Technique: LoRA Fine-Tuning **Experiment:** 4 Models Trained (512² vs 1024² px)

Image Generation
Workflow: Custom ComfyUI Nodes
Enhancements: VAE
Output: "Mensa Food" Style Images

Quantitative Evaluation
Tools: FID, DINOv2, dgm-eval
Goal: Quantify Parameter Impact

Parameter Comparison Table

Parameter	v1	v2	v3	v4
Dataset Configuration				
Resolution	512	1024	1024	1024
Num Repeats	20	20	30	40
Flip Augmentation	false	true	false	false
Color Augmentation	false	true	false	false
Training Schedule				
Max Train Epochs	20	10	10	10

Quantitative Results

Evaluation

To assess image quality and diversity, we used dgm-eval [4] to compare generations from the base model and two LoRA fine-tuned models. Metrics such as Fréchet Distance (FD), precision, recall, density, and coverage were evaluated against a reference dataset of real Mensa food images.

Real image 48 of the dataset

Both LoRA models outperformed the base model. FD decreased notably, indicating generated distributions were closer to real data. Precision and recall improved, suggesting higher visual fidelity and better diversity. LoRA Model 2 achieved the lowest FD and highest precision, while Model 1 showed slightly better density and coverage.

To test generalization, we evaluated the models on previously unseen meals. Results showed an FD of 1784.36, precision of 0.88, and perfect recall (1.0), indicating good distributional coverage but slightly reduced visual fidelity. Lower density (0.528) and coverage (0.72) suggest weaker local consistency and spread compared to the training set.

Notably, FD is test set-dependent—it reflects the distance between generated and real distributions from a specific reference set. Thus, higher FD on new meals likely reflects a distributional shift rather than a drop in generation quality, highlighting limited generalization beyond the fine-tuning set.

Overall, fine-tuning with just 50 images significantly improved realism and diversity in generated Mensa food images, though generalization remains limited.

No LoRA

FID	1639.27
FID (infinity value)	1543.90
precision	0.86
recall	0.96
density	0.58
coverage	0.9

LoRA v3

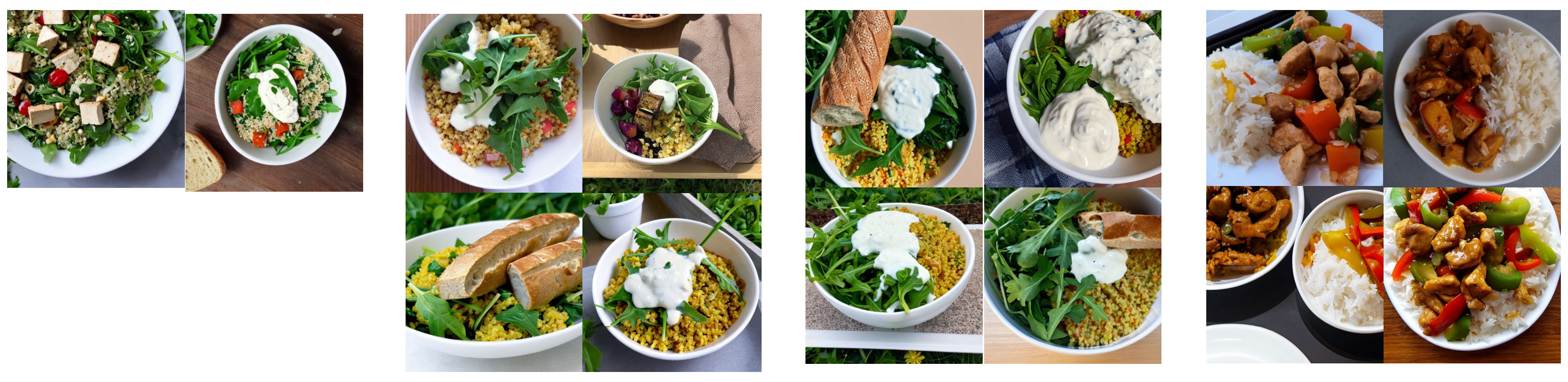
FID	1180.56
FID (infinity value)	1096.151
precision	0.96
recall	1.0
density	0.78
coverage	0.98

LoRA v4

FID	1111.77
FID (infinity value)	1007.86
precision	0.98
recall	1.0
density	0.72
coverage	0.94

Unseen

FID	1784.36
FID (infinity value)	1714.285
precision	0.88
recall	1.0
density	0.53
coverage	0.72



Generated images with each model for prompt:
mensa-food, a high-angle view of a white bowl with bulgur salad, organic tofu cubes, leaf spinach, arugula, and a dollop of white sauce, a piece of baguette on the side, on a white tray, outdoors, bright lighting.
The unseen images are generated with the prompt:
a white plate with stir-fried chicken, bell peppers, and a side of steamed jasmine rice, bright lighting.

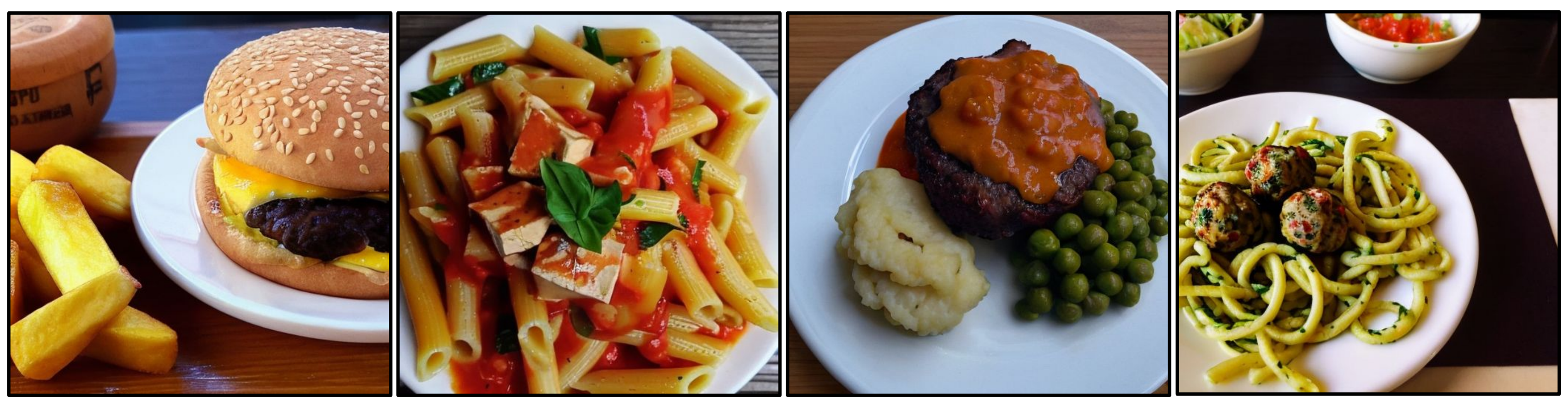
Qualitative Results

Visually, the models fine-tuned with LoRA show a dramatic improvement over the base Stable Diffusion model. Both LoRA v3 and v4 produced the most coherent and stylistically accurate images, successfully capturing the unique, non-professional aesthetic of the Mensa food dataset. While large-scale models like DALL-E 3 or Gemini tend to generate idealized, "fancy" food photography, our fine-tuned model excelled at its intended purpose: replicating the authentic, real-world look of the training images.

The process itself was highly accessible. Generating images was straightforward using a ComfyUI workflow, and the training was made possible by leveraging Google Colab and the wealth of available online documentation for LoRA.

However, the primary limitation of our approach stems from the "garbage in, garbage out" principle. The quality of the generated images is fundamentally tied to the quality of the training data. Our dataset, while authentic, consisted of images with poor lighting and inconsistent camera angles, which are reflected in the final outputs.

For future work, the most impactful improvement would be to curate a higher-quality dataset. By generating our own images with controlled lighting, a better camera, and a wider variety of angles, we could significantly enhance the realism and fidelity of the results. Nevertheless, this project successfully proves that with even a small, imperfect dataset, LoRA provides a highly effective and feasible path to adapt a powerful foundation model to a niche visual domain.



References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *arXiv.org*, Dec. 20, 2021. https://arxiv.org/abs/2112.10752
[2] E. J. Hu *et al.*, "LORA: Low-Rank adaptation of Large Language Models," *arXiv.org*, Jun. 17, 2021. https://arxiv.org/abs/2106.09685
[3] Hollowstrawberry, "GitHub - hollowstrawberry/kohya-colab: Accessible Google Colab notebooks for Stable Diffusion Lora training, based on the work of kohya-ss and Linaqruf," *GitHub*. https://github.com/hollowstrawberry/kohya-colab
[4] G. Stein et al., "Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models" Advances in Neural Information Processing Systems, vol. 36, 2023.

Acknowledgement: