Enhancing Text-To-Image Generation with LLMs

Aakrist Godar Harshith Mulakala Jagadeep Kalluri Sean Varghese Swayam Arora
Arjun Junghare Dr. Xinya Du

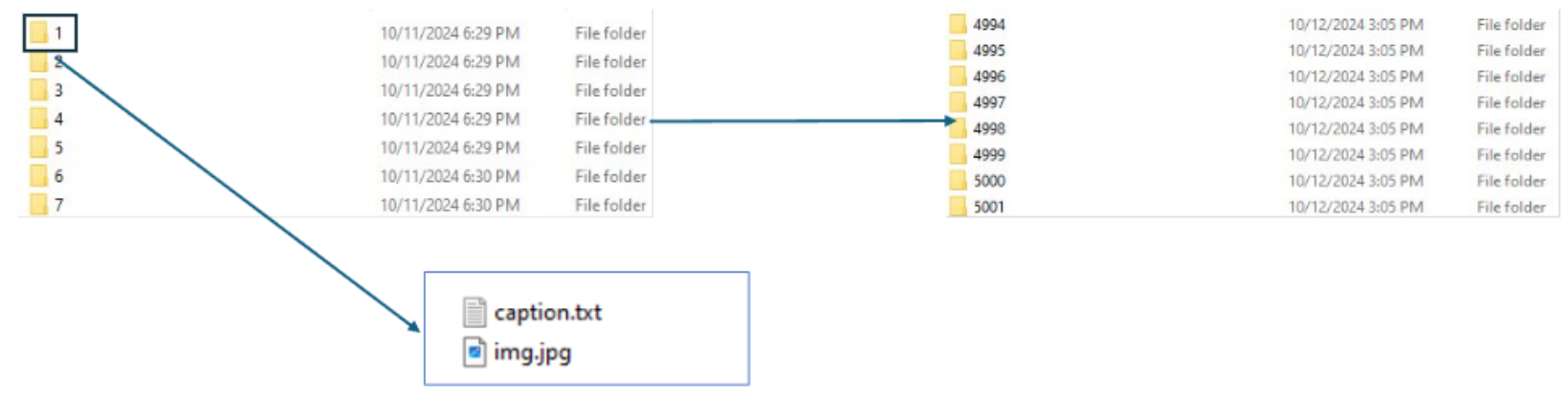
Introduction

Text-to-image generation has advanced with the help of large language models (LLMs), but challenges in achieving specificity and contextual accuracy remain. This project aims to enhance existing LLM-based image generation models to improve their ability to produce images that are both accurate and contextually relevant. By fine-tuning these models with specialized datasets, we seek to bridge the gap between general image generation and the need for precise, tailored outputs. This work provides a foundation for more responsive and context-aware visual generation, evaluated through comparative analysis of image quality and contextual alignment.

Dataset

PET BONE BROTH
Helps nourish your dog's skin, coat, joints, digestive system, hydration, and overall health!
Just add to warm water!
PET BONE BROTH

Figure 1. Sample Ad



1	10/11/2024 6:29 PM	File folder
2	10/11/2024 6:29 PM	File folder
3	10/11/2024 6:29 PM	File folder
4	10/11/2024 6:29 PM	File folder
5	10/11/2024 6:29 PM	File folder
6	10/11/2024 6:30 PM	File folder
7	10/11/2024 6:30 PM	File folder

4994	10/12/2024 3:05 PM	File folder
4995	10/12/2024 3:05 PM	File folder
4996	10/12/2024 3:05 PM	File folder
4997	10/12/2024 3:05 PM	File folder
4998	10/12/2024 3:05 PM	File folder
4999	10/12/2024 3:05 PM	File folder
5000	10/12/2024 3:05 PM	File folder
5001	10/12/2024 3:05 PM	File folder

caption.txt
img.jpg

Figure 2. Caption and Image labelling

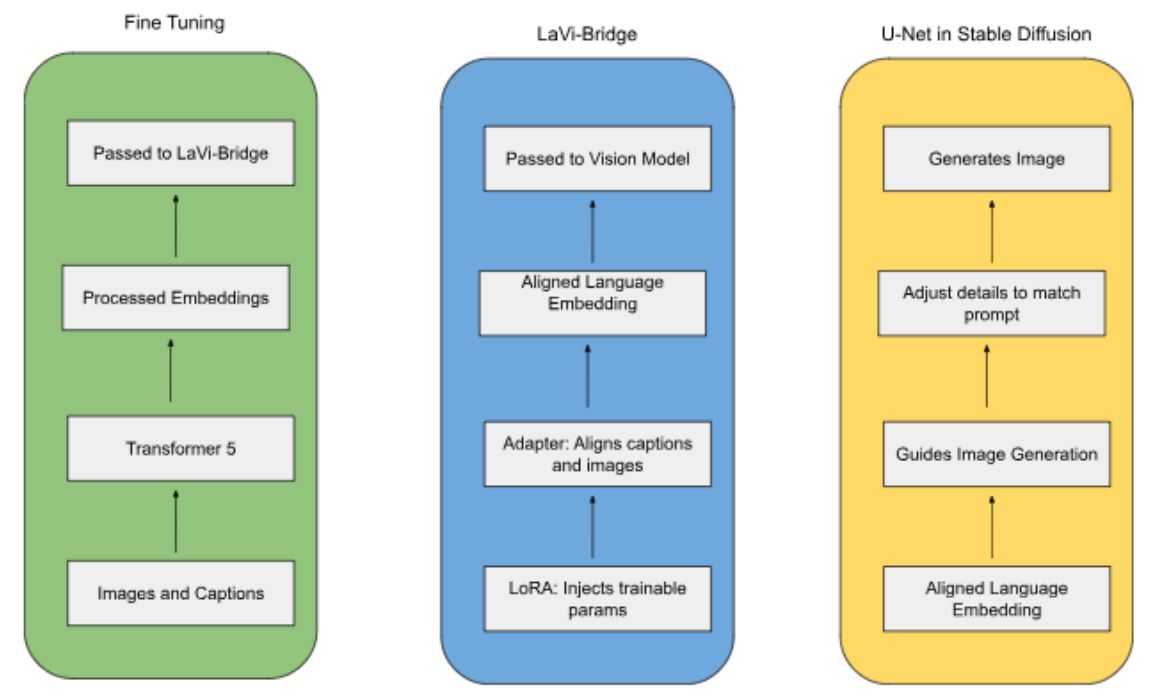
- Dataset sourced from the Meta Ads Library, a public database of ads from Meta platforms (e.g., Facebook, Instagram, WhatsApp).
- Used Selenium to search for 200 expansive keywords and scrape 25 unique ads per keyword.
- Collected both images and their respective captions for each ad for each keyword
- Final dataset organized into 5,000 folders labeled sequentially (1 to 5000).
- Each folder contains: An image file named `img.jpg` (the ad image) & A caption file named `caption.txt` (the ad caption).

Method

LaVi-Bridge is a flexible pipeline that connects pre-trained language and vision models. It retains the original weights of both models, modifying them minimally. The system incorporates Low-Rank Adaptation (LoRa) and Adapters to improve compatibility and customization. LoRa introduces trainable parameters into both models, allowing fine-tuning with custom data while preserving the original weights. Adapters act as a bridge between the language and vision models, enhancing the ability of LaVi-Bridge to process and understand text more effectively than traditional models. The system includes two key enhancements: a fine-tuned T5 language model trained on scraped captions for two epochs with a learning rate of 1 * 10-̂4, optimizing its text-to-image alignment capabilities, and a comprehensively trained LaVi-Bridge component that processes both images and captions simultaneously. This dual-modal training approach enables more sophisticated associations between textual descriptions and visual elements, resulting in improved integration and control over the image generation process.

Experiment

- Fine-tuned T5 on captions for improved text-to-image alignment.
- Trained for 2 epochs at a learning rate of 1×10^{-4}.
- Goal: Enhance T5's contextually accurate prompt generation.
- LaVi-Bridge fine-tuned on images and captions for better matching.



Fine Tuning
Images and Captions
Transformer 5
Processed Embeddings
Passed to LaVi-Bridge

LaVi-Bridge
LoRA: Injects trainable params
Adapter: Aligns captions and images
Aligned Language Embedding
Passed to Vision Model

U-Net in Stable Diffusion
Aligned Language Embedding
Guides Image Generation
Adjust details to match prompt
Generates Image

Figure 3. Image comparisons between models

Results

Figure 4 demonstrates outputs from three model configurations (T5, T5+Lavi Bridge, Fine-tuned T5+Lavi Bridge) across six distinct prompts. The prompts encompass: (1) a red sports car against a coastal sunset; (2) a joyful dog with food bowl in a garden setting; (3) a model wearing winter attire in a snowy cityscape; (4) a tropical beach scene featuring a cabana; (5) an organic product arrangement with herbs and citrus; and (6) a luxury timepiece on mahogany. Each row represents a progressive model iteration, with the baseline T5 outputs in the top row, T5+Lavi Bridge in the middle, and Fine-tuned T5+Lavi Bridge outputs in the bottom row. Visual analysis indicates progressive enhancement in image quality, prompt adherence, and detail rendering across the model iterations.

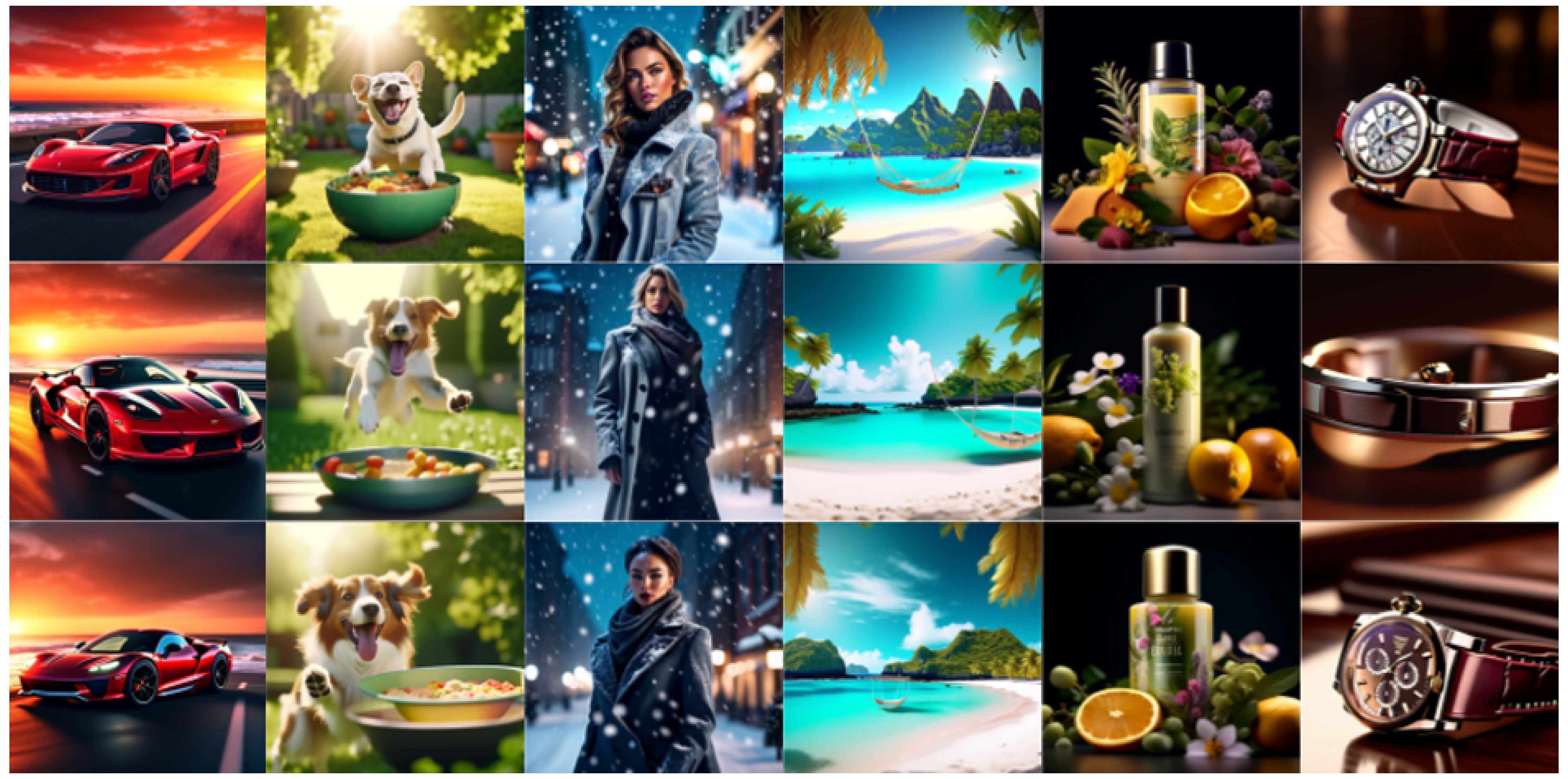


Figure 4. Image Comparison Matrix

Analysis

Model	Avg. CLIP Score
LaVi-Bridge (baseline)	0.344
Fine-tuned T5 + LaVi-Bridge (ours)	0.348
Fine-tuned T5 + Fine-tuned LaVi-Bridge (ours)	0.371

Table 1. Average CLIP scores for evaluated models

CLIP scores determine how close an image is aligned with a text description making it an effective indicator of how well an image can be generated. Our models demonstrate how a fine-tuned T5 model, trained on Meta's Ads Library data, enhances text-to-image generation using the LaVi Bridge architecture. By leveraging LoRa adapters and our specialized training approach, we achieved a 6.4 % improvement in CLIP scores over baseline models. Our model excels particularly in capturing specific details, showing marked improvement in generating contextually appropriate visuals for marketing content. Though, this approach faces limitations in both training and evaluation. Human benchmarking can reveal subjective variations in ad effectiveness assessment, particularly for complex ad narratives that require understanding of brand voice or marketing strategy. Furthermore, certain nuanced details in ad copy may not be accurately reflected in generated visuals, especially for niche market segments underrepresented in our training data. The model occasionally struggles with complex multi-product advertisements or highly specialized industry terminology, suggesting areas for future dataset expansion.

Conclusion

Our AdFusion implementation demonstrates that fine-tuning T5 with LaVi Bridge significantly improves text-to-image ad generation, achieving a CLIP score of 0.3703 compared to the baseline's 0.348. This enhancement creates more contextually accurate and visually aligned advertising content. Future work could focus on: (1) expanding the dataset beyond Meta's Ad Library, (2) increasing training epochs while maintaining efficiency, (3) incorporating multimodal transformers for better text-visual alignment, and (4) developing more robust evaluation metrics beyond CLIP scores. These improvements would further advance the model's capability to generate highly relevant and engaging advertising visuals.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv*. `https://doi.org/10.48550/arXiv.1706.03762`

[2] Zhao, S., Hao, S., Zi, B., Xu, H., & Wong, K.-Y. K. (2024). Bridging Different Language Models and Generative Vision Models for Text-to-Image Generation. *arXiv*. `https://doi.org/10.48550/arXiv.2403.07860`

"Enhancing Text-To-Image Generation with LLM " is published by ACM Research, a registered student organization. "Enhancing Text-To-Image Generation with LLMs" is not an official publication of UT Dallas and does not represent the views of the university or its officers. The University of Texas at Dallas is an Equal Opportunity/Affirmative Action University. Students with disabilities needing special assistance to attend please call (972-883-2946) [or the number of Fraternity and Sorority Life (972-883-6523)]. Texas Relay Operation: 1-800-RELAYTX.

ACM Research Fall 2024