The Minimalist Revolution5: Democratizing High-Quality Design with Efficient AI

Paul Hornig

Abstract

This paper explores the resource-efficient fine-tuning of a Stable Diffusion model for the generation of minimalist logos. By leveraging a multimodal approach that combines text prompts with structural sketches via ControlNet, we address the limitations of standard text-to-image models in adhering to specific design constraints. We fine-tune the model on a curated dataset of 1,500 minimalist logos using consumer-grade hardware. A core contribution of this work is a systematic analysis of Low-Rank Adaptation (LoRA) hyperparameters, identifying optimal configurations for stylistic coherence and structural fidelity. We present a quantitative evaluation using CLIP, FID, and SSIM metrics, alongside qualitative case studies, demonstrating that high-quality, controllable design automation is achievable with limited computational resources.

Contents

1 Introduction

Minimalist logo design relies on reduction, clarity, and structural precision - qualities that are often challenging for general purpose generative models to achieve consistently. While large-scale text-to-image models excel at artistic composition, they frequently struggle to produce the clean, vector-like aesthetics required for professional branding or to strictly adhere to a user's layout constraints [1, p. 1].

This work presents a specialized pipeline for generating minimalist logos by fine-tuning Stable Diffusion v1.5 [2]. To ensure both semantic relevance and structural control, we employ a hybrid conditioning strategy: text prompts define the style and content, while ControlNet [3][4, p. 8][5, p. 5] enforces the geometric structure based on input sketches. Unlike approaches requiring massive datasets and industrial compute clusters, we focus on resource efficiency. We demonstrate that a compact, high-quality dataset of 1,500 examples is sufficient to adapt the model to the minimalist domain using consumer-grade hardware (NVIDIA RTX 5080).

Our research focuses on the optimization of the fine-tuning process itself. We conduct a rigorous analysis of Hyperparameters within the Low-Rank Adaptation (LoRA) technique, specifically examining the impact of learning rates and rank dimensions on the trade-off between training stability and generation quality.

The evaluation is primarily quantitative, utilizing the Fréchet Inception Distance (FID) to assess image quality, the CLIP score for semantic alignment, and the Structural Similarity Index (SSIM) to measure fidelity to the input sketches. We complement these metrics with qualitative case studies that illustrate the model's capability to translate rough sketches into polished, minimalist designs. This approach validates that accessible hardware and efficient training strategies can yield professional-grade design automation tools.

Hypotheses

Based on the defined objectives, we postulate the following hypotheses:

- **Hypothesis 1 (H1):** Additional conditioning of the model via a sketch leads to a significantly higher structural correspondence with the design intent compared to purely text-conditioned results.
- **Hypothesis 2 (H2):** Optimizing an image generator through fine-tuning on a specialized logo dataset significantly improves the ability to generate stylistically coherent logos, as reflected in established image quality metrics.
- **Hypothesis 3 (H3):** Specific hyperparameter configurations have a direct and measurable influence on result quality, with an optimal configuration leading to better visual quality, structural correspondence, and text-image coherence.
- **Hypothesis 4 (H4):** A resource-efficiently optimized prototype generates results on consumer-grade hardware that exhibit higher quality in realizing design intent and greater commercial relevance in a qualitative evaluation compared to the unspecialized base model.

2 Theoretical Background and Related Work

2.1 Minimalist Logo Design

Principles and Criteria

The design philosophy of minimalism is scientifically grounded in Ockham's Razor, which posits that among functionally equivalent alternatives, the simplest is preferable [6, p. 172]. In the context of logo design, minimalism aims to reduce a brand identity to its essential elements to achieve maximum **clarity**, recognizability, and functionality. A minimalist logo eliminates superfluous details and complex structures in favor of simple shapes, clear lines, and a limited color palette [7, p. 52]. This reduction enhances **memorability**, as unique, simple forms are easier to store in visual memory [8, p. 2]. Furthermore, it promotes **timelessness** by avoiding trends [7, p. 40] and ensures **versatility**, allowing the design to remain recognizable across all scales and media [7, p. 44].

Topology of Minimalist Logos

Following Wheeler [7, p. 51], minimalist logos can be classified into three main categories:

- **Typographic Logos:** Wordmarks (e.g., Google) or Monograms (e.g., IBM) that rely on type and negative space [9, p. 68].
- **Pictorial Marks:** Symbols representing the brand. These include Emblems (text inside symbol), Pictorial Marks (literal representations like Apple), and Abstract Marks (geometric forms like Nike) [10].

- **Combination Marks:** Integrating text and symbol (e.g., PayPal) to reinforce the brand message [10].

2.2 Generative AI Models

Generative AI has evolved significantly from Generative Adversarial Networks (GANs) [11], which often suffered from training instability [12, pp. 2–3]. Denoising Diffusion Probabilistic Models (DDPMs) [13, p. 3] introduced a more stable paradigm based on reversing a noise addition process. Rombach, Blattmann, Lorenz, *et al.* [14] further advanced this with Latent Diffusion Models (LDMs), which operate in a compressed latent space. This approach, used in Stable Diffusion, drastically reduces computational requirements, enabling high-resolution generation on consumer hardware.

2.3 Conditioned Modeling

To control generation, models combine modalities, typically text and image [3].

CLIP: Bridging Text and Image

Radford, Kim, Hallacy, *et al.* [15] introduced CLIP (Contrastive Language-Image Pre-training), which learns a joint embedding space for text and images. This allows diffusion models to be guided by text prompts, aligning the generated image with semantic descriptions [4].

ControlNet: Structural Control

While CLIP provides semantic control, it lacks spatial precision. ControlNet [3] addresses this by adding a trainable copy of the model's encoder blocks. This allows the injection of structural conditions, such as edge maps or sketches, directly into the generation process without retraining the base model. This is crucial for logo design, where specific shapes must be preserved.

2.4 Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning large foundation models is resource-intensive. PEFT methods aim to adapt models by training only a small subset of parameters [16, p. 2]. **Low-Rank Adaptation (LoRA)** [17, p. 4] is a prominent PEFT technique. It hypothesizes that weight updates have a low intrinsic rank and approximates them using low-rank matrices ($\Delta W = B \cdot A$). This significantly reduces the number of trainable parameters and memory usage, making fine-tuning feasible on consumer GPUs [18, p. 4].

2.5 Evaluation Metrics

We employ three metrics to evaluate the generated logos:

- **CLIP Score:** Measures the semantic alignment between the generated image and the text prompt [19].

- **SSIM (Structural Similarity Index):** Quantifies the structural fidelity of the generated logo to the input sketch, focusing on luminance, contrast, and structure [20][21, p. 3].
- **FID (Fréchet Inception Distance):** Assesses the realism and quality of generated images by measuring the distance between the feature distributions of real and generated images [22, p. 6].

3 Methodology

3.1 Research Design and Dataset

This study employs a quantitative-experimental approach to evaluate the performance of fine-tuned diffusion models for minimalist logo generation. The core objective is to adapt a base model using parameter-efficient fine-tuning (PEFT) and structural control techniques.

The foundation for training is the `iamkaikai/amazing_logos_v4` dataset [23], a multimodal corpus of approximately 400,000 logo-caption pairs. This dataset was chosen for its scale and detailed metadata, which includes company names, descriptions, and stylistic tags.

3.2 Model Selection and Architecture

Base Model

Stable Diffusion v1.5 [2] was selected as the base model. Its open-source nature, extensive ecosystem, and balance between generative quality and computational requirements make it ideal for iterative optimization on consumer hardware (NVIDIA RTX 5080, 16GB VRAM).

Structural Control: The ControlNet Pipeline

To enforce structural fidelity to user sketches, we utilize ControlNet [3]. We evaluated three pretrained models: Canny, Scribble, and Lineart.

- **Canny:** Often resulted in rigid, uncreative outputs due to strict edge adherence.
- **Scribble:** Too loose, often ignoring fine structural details of the input sketch.
- **Lineart:** Identified as the **optimal solution**. It respects the precise form of the sketch while allowing the diffusion model sufficient freedom to interpret style and texture (Figure 1).

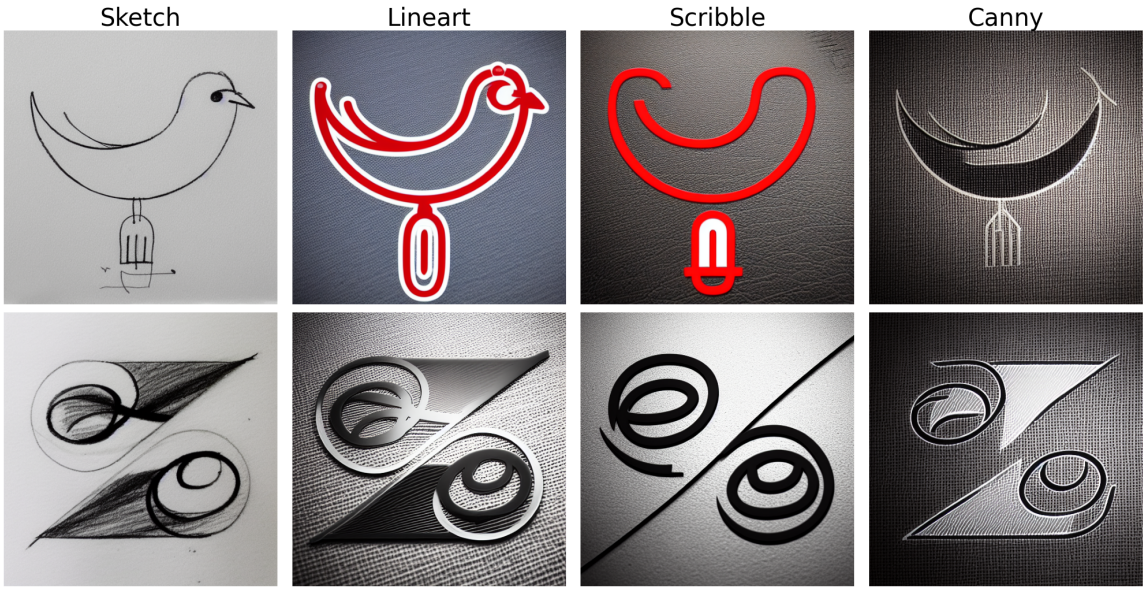


Figure 1: Empirical comparison of ControlNet models for conditioned logo generation.

3.3 Data Preparation

A multi-stage pipeline was implemented to curate a high-quality subset for training.

Metadata Cleaning and Enrichment

The raw dataset contained semi-structured captions that required systematic cleaning to ensure high-quality semantic control. The processing pipeline involved three key steps:

1. **Normalization and Parsing:** Raw text strings were cleaned of special characters and split into structured fields: `company`, `description`, `category`, and `tags`.
2. **Heuristic Correction:** Inconsistencies were addressed using heuristic rules. For instance, misplaced tags were reassigned, and swapped content between `description` and `category` was corrected based on string length analysis.
3. **Category Consolidation:** To enable effective classification, the number of categories was reduced from 44,810 to 109 by grouping synonyms (e.g., merging "tech_startup" into "technology") and removing rare instances. These were further aggregated into 10 top-level categories.

Table 1 illustrates the transformation from the raw caption to the structured metadata.

Table 1: Comparison of data structure before and after cleaning

Before Cleaning	After Cleaning
caption: Simple elegant logo for Aziz Firaat, Read Book Tick Checkmark Todo List, Website, successful vibe, minimalist, thought-provoking, abstract, recognizable, relatable, sharp, vector art, even edges	**company:** Simple elegant logo for Aziz Firaat
	description: Read Book Tick Checkmark Todo List
	tags: successful_vibe, minimalist, thoughtprovoking, abstract, recognizable, relatable, sharp, vector_art, even_edges
	category_main: tech
	category: web_digital

Curation and Minimalism Score

To ensure the dataset reflects the target aesthetic, we developed a weighted "Minimalism Score" (0-100) based on features such as color count, edge density, contour complexity, and whitespace ratio.

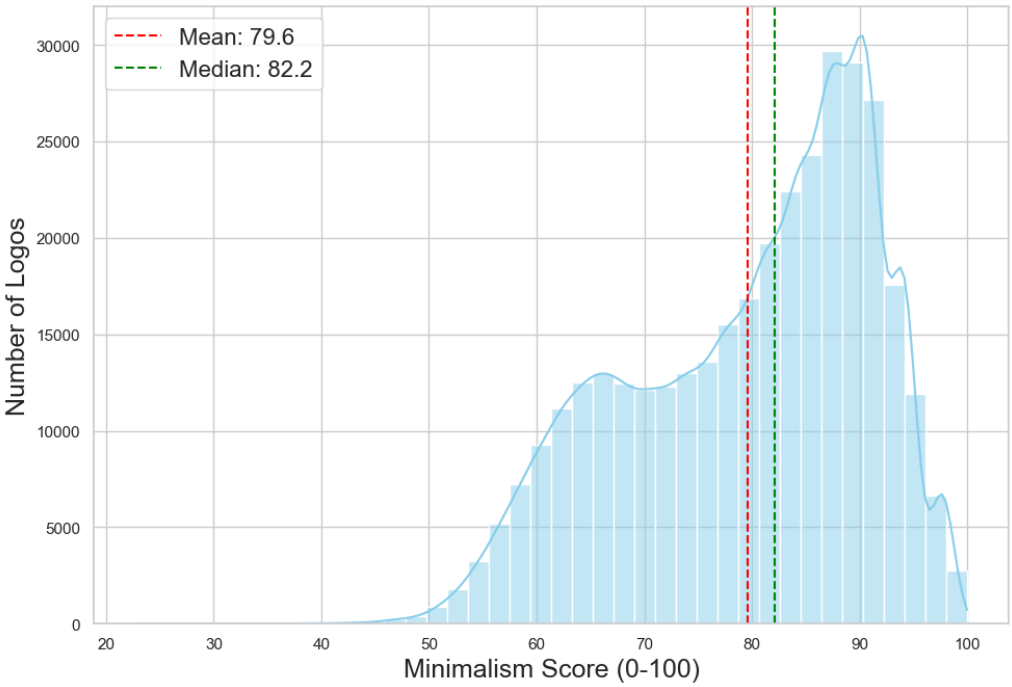


Figure 2: Distribution of the calculated Minimalism Score in the dataset.

As shown in Figure 2, we filtered the dataset using the median score (82.2) as a threshold. From the remaining pool, a class-balanced subset of 1,810 images was extracted to facilitate efficient hyperparameter optimization [3], [24].

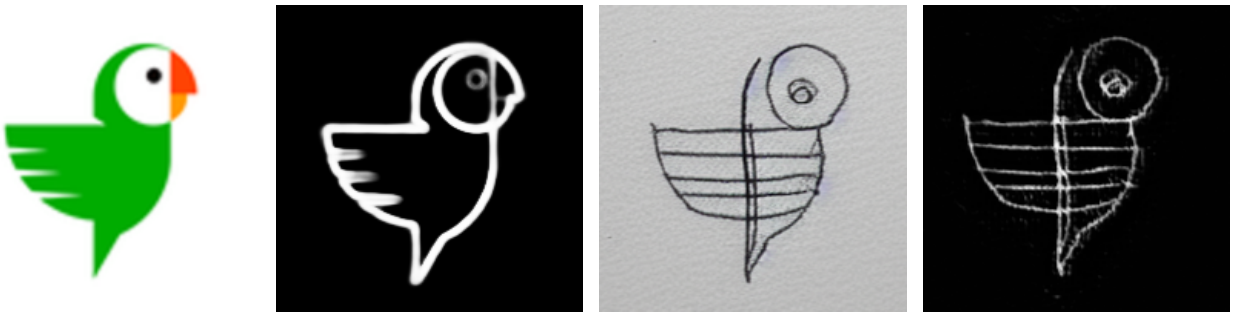
Generation of Control Signals

Since the dataset contains finished logos, synthetic "sketches" were generated to train the model's response to rough inputs.

1. **Sketch Generation:** We used Stable Diffusion v1.5 with the Scribble ControlNet to transform original logos into abstract, hand-drawn-style sketches. The Scribble model was chosen here to simulate the imperfection of human drawings.

2. **Lineart Map:** These sketches were then processed into clean Lineart maps to serve as the precise control signal for the training pipeline.

The process is illustrated in Figure 3.



(a) Original Logo **(b)** Scribble Map **(c)** Generated Sketch **(d)** Lineart Map

Figure 3: Generation of synthetic control signals (sketches) from the dataset.

Prompt Engineering and Data Split

Text prompts were standardized to the format: `minimalistic logo, solid background; description: {description}; tags: {tags}`. The curated dataset (1,810 images) was split into **Training (80%, 1448)**, **Validation (10%, 181)**, and **Test (10%, 181)** sets using a fixed seed.

3.4 Fine-Tuning Strategy and Hyperparameters

We employ Low-Rank Adaptation (LoRA) [17] to fine-tune the UNet component of Stable Diffusion. The VAE, Text Encoder, and ControlNet are frozen to preserve their pretrained capabilities and ensure resource efficiency.

We systematically analyze the following hyperparameters to identify the optimal configuration:

- **Target Modules:** Comparing adaptation of only Attention layers vs. an Extended configuration (Attention + Feed-Forward layers) [25].

- **LoRA Rank (r):** Evaluating ranks **{4, 8, 16, 32}** to find the balance between expressivity and efficiency [17].

- **LoRA Alpha (α):** Fixed at $1.5 \times r$. This decision is based on the finding of Hu, Shen, Wallis, *et al.* [17, p. 4] that the effective strength of LoRA adaptation can be primarily controlled via the learning rate, making a separate, detailed tuning of α less critical.
- **Learning Rate:** Exploring the range 1×10^{-6} **to** 1×10^{-4}, as PEFT methods typically tolerate higher rates than full fine-tuning [16].
- **Batch Size:** Fixed at **8** to maximize hardware utilization.

4 Implementation

4.1 Environment

To ensure reproducibility and efficiency, a standardized environment was established.

- **Hardware:** NVIDIA RTX 5080 (16 GB VRAM) for mixed-precision training (FP16).
- **Software:** Python-based stack utilizing `torch` and `torchvision` for deep learning operations.
- **Libraries:** `diffusers` for the ControlNet pipeline, `transformers` for tokenization, and `peft` for efficient LoRA fine-tuning.
- **Tracking:** MLflow [26] containerized via Docker for experiment tracking and model management.

4.2 Pipeline Design

The implementation follows a modular pipeline approach separating training and evaluation to ensure independent validation. A fixed random seed of 42 is used throughout to eliminate stochastic variance.

Training Pipeline

The training process orchestrates the fine-tuning of Stable Diffusion v1.5 using LoRA and ControlNet.

1. **Initialization:** The base model, ControlNet (Lineart), and tokenizer are loaded in `bfloat16`. LoRA adapters are injected into the UNet, while the base model remains frozen.
2. **Training Loop:** For each step, images are encoded into latent space and noised. Text prompts are tokenized, and the model predicts noise based on the latents, text embeddings, and control maps.
3. **Optimization:** Gradients are calculated only for LoRA adapters using the MSE loss between predicted and actual noise. The AdamW optimizer updates the weights.

The training process is interrupted at regular intervals to perform validation on a separate dataset, ensuring monitoring of model progress and early detection of overfitting. Each experiment is trained for a fixed duration of 14 epochs, which corresponds to exactly 2,534 training steps (181 per epoch) given a batch size of 8 and 1,448 training images. This duration aligns with the guideline of 2,500 steps (at batch size 1) recommended by Cloneofsimo (GitHub Author) [27] for high-quality training and is consistent with findings by Ruiz, Li, Jampani, *et al.* [24], who achieved good results with as few as 1,000 iterations. The checkpoint saved after completing all 2,534 steps is used for final evaluation. All relevant hyperparameters and metrics are logged via MLflow.

Evaluation Pipeline

The evaluation pipeline quantifies model performance using the validation set.

- **CLIP Score:** Measures semantic alignment between the generated image and the text prompt [28].
- **SSIM:** Evaluates structural fidelity. To normalize the metric and focus solely on structure, the generated image is converted into a lineart map. SSIM is then computed between this extracted lineart and the original control map [29].
- **FID:** Assesses realism and diversity by comparing the distribution of generated images against the real dataset [30].

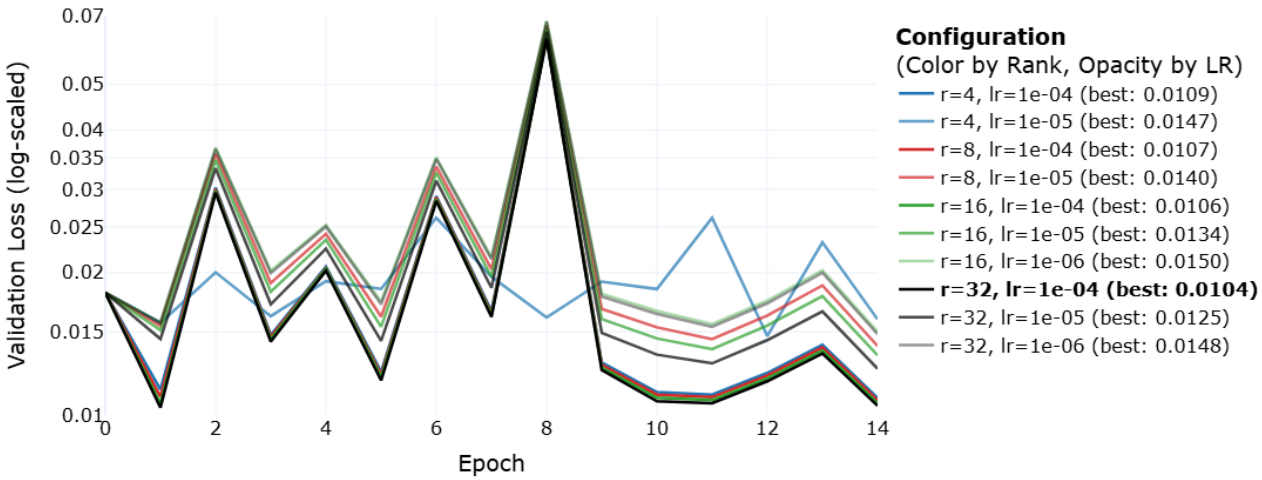
5 Results

5.1 Quantitative Evaluation

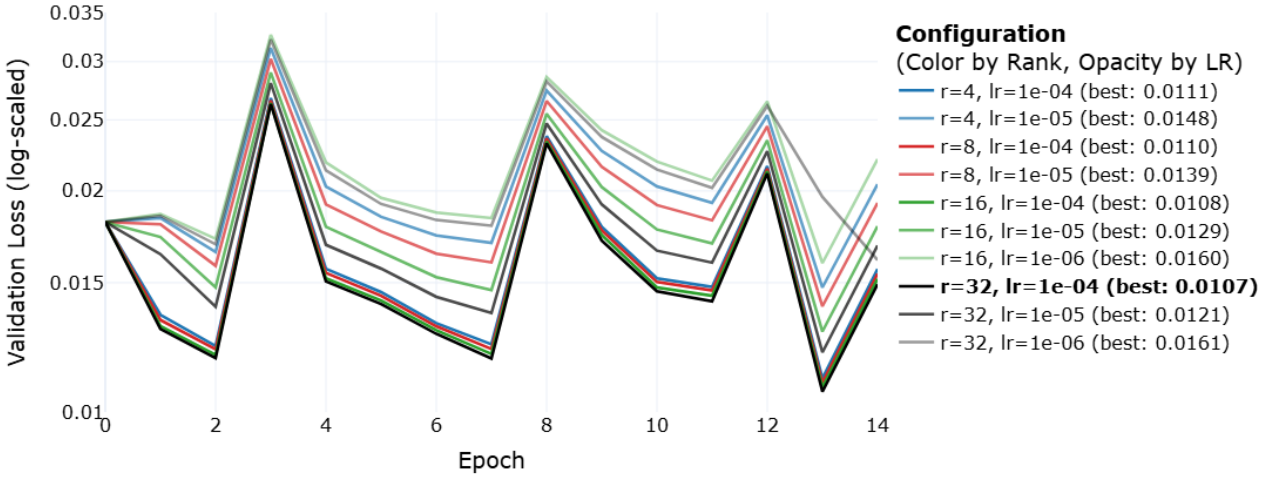
The quantitative evaluation objectively assesses the model's performance using CLIP score, SSIM, and FID, as defined in Chapter 4.2. All experiments were tracked via MLflow. The study focuses on the sensitivity of the model to LoRA rank (r), learning rate (lr), and the choice of adapted modules ("attn_only" vs. "extended").

Loss Curve Analysis

Validation loss serves as an indicator of generalization. Figures 4a and 4b illustrate the training dynamics.



(a) Configuration "attn_only"



(b) Configuration "extended"

Figure 4: Comparison of validation loss curves for "attn_only" and "extended" configurations across hyperparameters.

Observations for "attn_only" The learning rate is the primary driver, with $lr = 1e-4$ yielding the lowest loss. The LoRA rank has a minor impact, mostly noticeable at $lr = 1e-5$ where higher ranks perform slightly better. A notable exception is the combination of $r = 4$ and $lr = 1e-5$, which exhibits remarkably low variance, indicating a very stable learning process despite not achieving the

absolute lowest loss.

Observations for "extended" The learning rate is even more dominant here, with significant gaps between $lr = 1e-4$ and lower rates. While rank influence is generally low at the optimal learning rate, the combination of high rank ($r = 32$) and lowest learning rate ($lr = 1e-6$) shows an interesting anomaly: it achieves lower loss than the $lr = 1e-5$ configuration towards the end of training, suggesting that lower learning rates might benefit from extended training durations.

Synthesis The analysis reveals that the learning rate is the dominant factor. For both configurations, $lr = 1e-4$ consistently yields the lowest loss. The "attn_only" models generally achieve lower loss levels with less variance compared to "extended" models, which exhibit higher sensitivity and fluctuations. This instability is a known phenomenon in LoRA fine-tuning of diffusion models [31].

Structural Fidelity (SSIM)

SSIM measures how well the generated logo adheres to the input sketch. Data analysis reveals a complex interaction between learning rate, LoRA rank, and target modules, rather than a single dominant factor. No single hyperparameter shows a consistently superior trend.

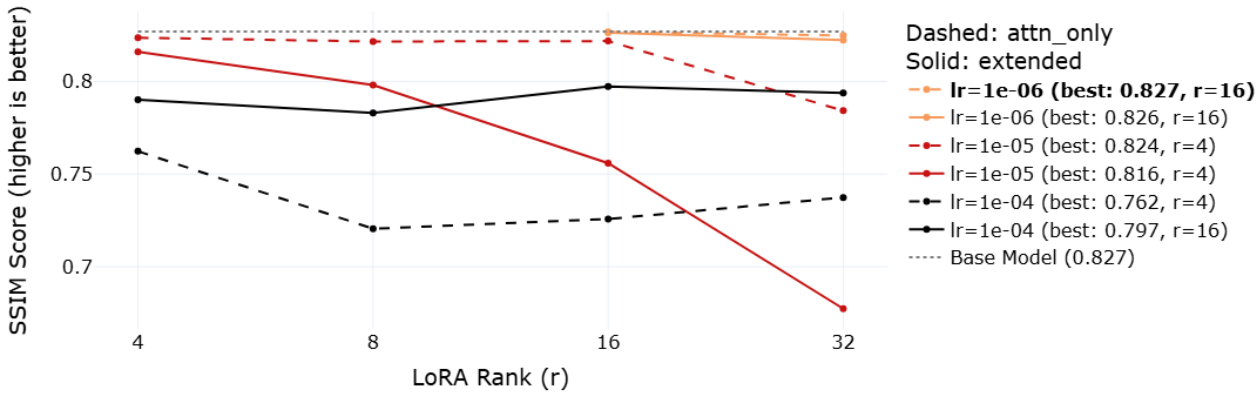


Figure 5: SSIM scores vs. learning rate, rank, and module configuration.

As shown in Figure 5, the pre-trained base model (dotted line) already achieves an excellent SSIM of 0.827. Fine-tuning does not significantly improve structural fidelity; in fact, higher learning rates ($1e-4$) in the "extended" configuration can slightly degrade it. This confirms that ControlNet alone provides robust structural control.

Semantic Alignment (CLIP Score)

The CLIP score evaluates the semantic correspondence between the image and the text prompt.

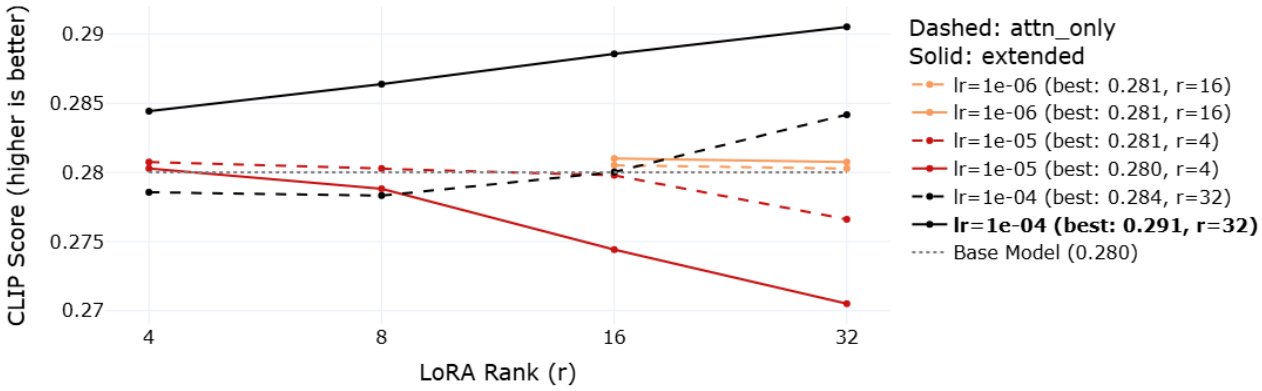


Figure 6: CLIP scores vs. learning rate, rank, and module configuration.

The analysis reveals complex interactions between the hyperparameters. Although the absolute differences in CLIP scores are small and within a small percentage range, clear trends can be identified: In contrast to SSIM, fine-tuning significantly improves semantic alignment (Figure 6). The best performance is achieved with $lr = 1e - 4$, high rank ($r = 32$), and the "extended" configuration, surpassing the base model by approximately 4%. This highlights the necessity of fine-tuning for capturing domain-specific semantics.

Image Quality (FID)

FID assesses the realism and feature distribution of the generated images.

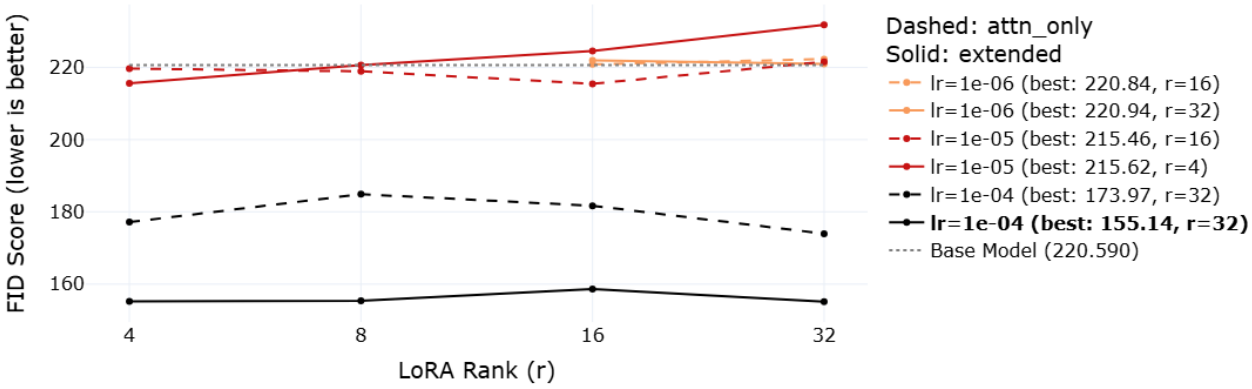


Figure 7: FID scores vs. learning rate, rank, and module configuration.

Figure 7 demonstrates that learning rate is the critical driver for image quality. The highest learning rate ($1e-4$) consistently produces the lowest (best) FID scores, improving upon the base model by roughly 28%. The "extended" configuration outperforms "attn_only" at this optimal learning rate.

Model Selection

To identify the optimal model, a combined metric normalizing SSIM, CLIP, and (inverted) FID was calculated.

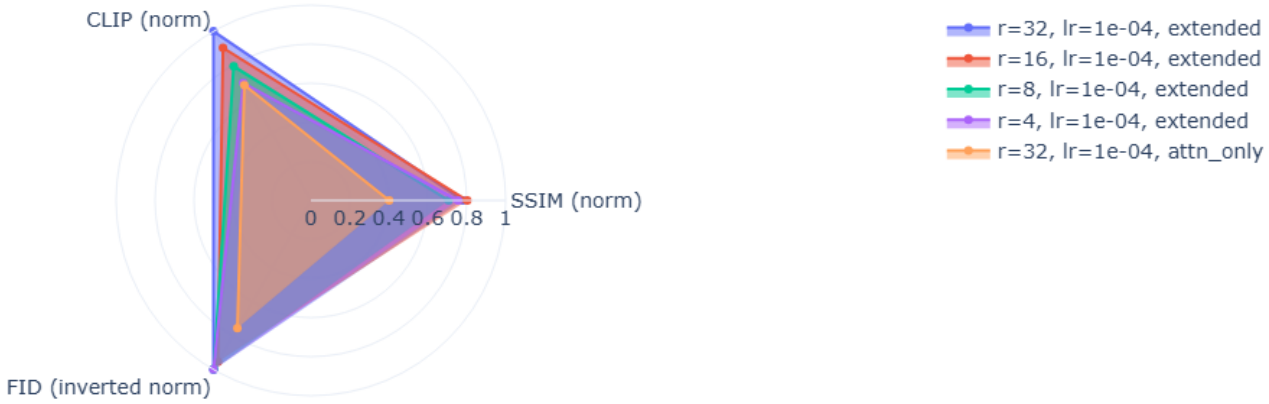


Figure 8: Top 5 model configurations (normalized metrics).

The radar chart in Figure 8 identifies the configuration with **Rank 32,** $lr = 1e-4$**, and "extended" modules** as the best overall performer. It offers the best compromise, maximizing image quality and semantic alignment while maintaining acceptable structural fidelity.

Evaluation on Test Set

The selected model was evaluated on an unseen test set to assess generalization. Table 2 compares the base model (without and with ControlNet) against the fine-tuned model.

Model	**CLIP Score ↑**	**FID ↓**	**SSIM ↑**
Base Model (w/o CN)	0.284	223.2	0.617
Base Model (w/ CN)	0.274	229.0	0.817
Finetuned Model	0.284	158.2	0.789
Improvement (Finetuned vs. Base w/ CN)	+3.7%	-30.9%	-3.4%

Table 2: Test set metrics comparison.

The results demonstrate the distinct roles of the components:

- **ControlNet** is essential for structure, boosting SSIM by 32.4% compared to the unconditioned base model.
- **Fine-tuning** restores semantic alignment (+3.7% CLIP) and drastically improves image quality (-30.9% FID). The significant FID reduction (229.0 to 158.2) outweighs the minor 3.4% SSIM decrease (0.817 to 0.789), representing a conscious trade-off: minimal structural precision is sacrificed for enhanced semantic and visual quality.

The fine-tuned model successfully combines structural control with domain-specific aesthetic quality.

5.2 Qualitative Analysis

To validate the quantitative findings, a qualitative analysis was conducted using five representative case examples (E1–E5), as illustrated in Figure 9. Four examples (E1, E2, E3, E5) were selected based on their sketches from the test dataset to cover diverse logo types: wordmarks, pictorial/abstract marks, and combined marks. Additionally, Example E4 - a monogram based on a hand-drawn sketch - was included to test the model on authentic human input. This diversity allows for a comprehensive evaluation across design categories.

For each case, both the base model (Stable Diffusion v1.5 with ControlNet) and the fine-tuned model generated a logo using the corresponding sketch and text prompt. To ensure objectivity and reproducibility, generation was performed with exactly one iteration per model, without any post-selection or optimization (no cherry-picking).

Negative Prompt (for all case examples): *sketch, photorealistic, pattern in background, noisy, blurry, watermark*

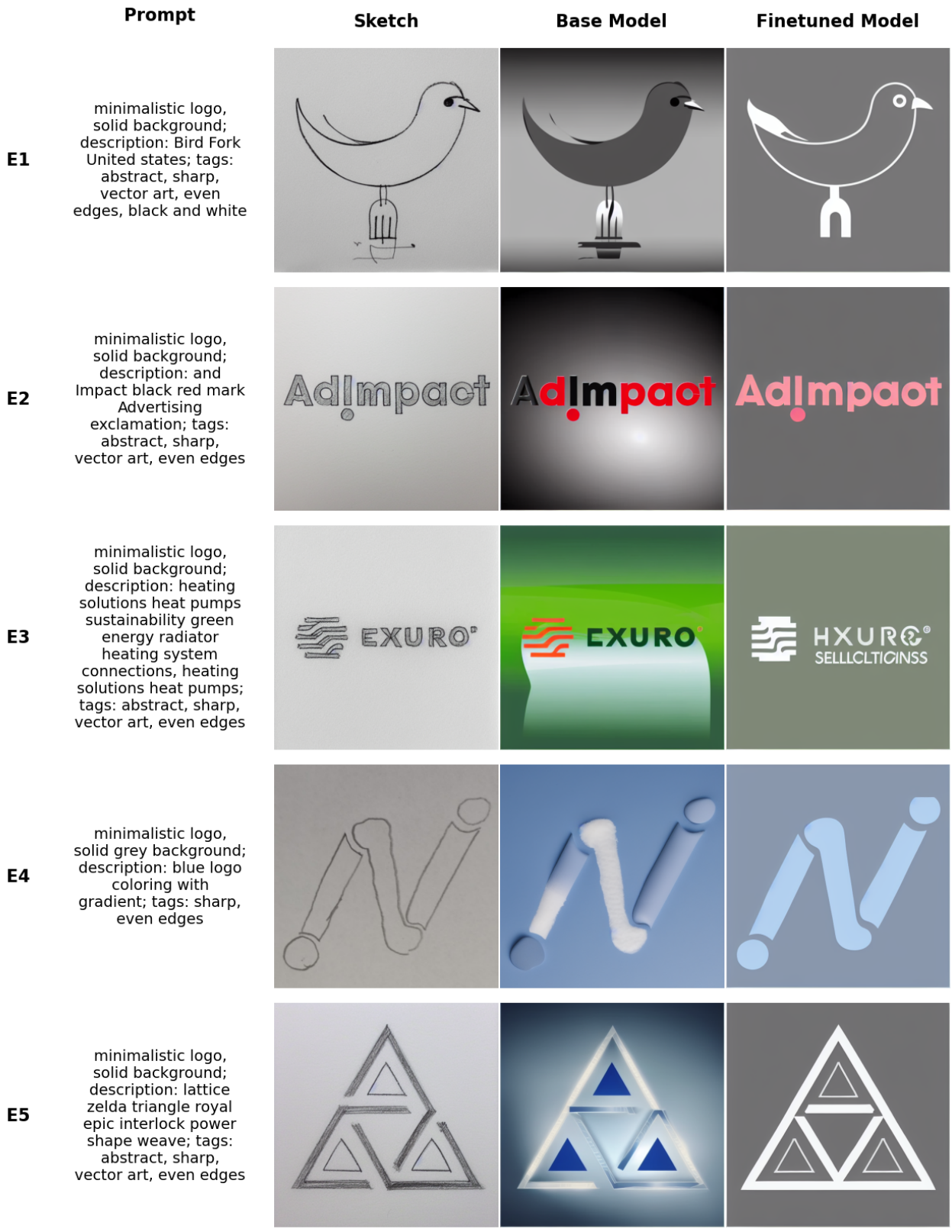


Figure 9: Comparison of generated logos across five representative cases (E1-E5) between the base and fine-tuned models.

Stylistic Evaluation

The fine-tuned model demonstrates a superior ability to implement specific design constraints compared to the base model.

- **Strengths:** It reliably generates "solid backgrounds" and consistent, minimalist color palettes, avoiding the unwanted textures often produced by the base model.
- **Weaknesses:** Challenges remain in fine-grained details. For instance, text rendering (Case E3) can be deformed, and specific gradient instructions (Case E4) are occasionally ignored.

Despite these limitations, the fine-tuned model shows a significant improvement in generating commercially viable, minimalist logos.

5.3 Confirmation of Hypotheses

Hypothesis H1, stating that sketch conditioning leads to significantly higher structural correspondence, was clearly confirmed. Evaluation showed a dramatic increase in the **SSIM!** (**SSIM!**) score from 0.617 for the base model without ControlNet to 0.817 with ControlNet (+32.4%). This validates the theoretical considerations of Zhang, Rao, and Agrawala [3], positioning ControlNet as a solution for the geometric limitations of pure text-to-image models.

Hypothesis H2, postulating quality improvement through domain-specific fine-tuning, was fully confirmed. The best fine-tuned model achieved a **CLIP!** (**CLIP!**) score of 0.284 vs. 0.274 for the base model (+3.7%) and an **FID!** (**FID!**) reduction from 229.0 to 158.2 (-30.9%). These results align with Ruiz, Li, Jampani, *et al.* [24], validating significant gains even with limited iterations. Notably, the slight SSIM drop (0.817 to 0.789) reflects a desired tradeoff: shifting focus from rigid structure to semantic accuracy and visual quality suitable for minimalist logos.

Hypothesis H3, assuming a correlation between hyperparameters and quality, was also confirmed. Learning rate proved to be the dominant factor for both semantic coherence (**CLIP!**) and visual quality (**FID!**). A high learning rate of $1e-4$ combined with the "extended" configuration yielded the best overall results. Higher **LoRA!** (**LoRA!**) ranks (32) tended to enable better semantic coherence at high learning rates, consistent with Hu, Shen, Wallis, *et al.* [17], suggesting that higher ranks approach full fine-tuning capacity.

Hypothesis H4, suggesting the superiority of the optimized prototype in qualitative assessment, was supported by the stylistic analysis. The evaluation (see Section 5.2) indicated that the fine-tuned model better adheres to the minimalist design intent compared to the base model. Specifically, it demonstrated improved capabilities in generating solid backgrounds and harmonic color palettes suitable for professional branding. While the base model frequently introduced unwanted textures, the optimized prototype produced cleaner, vector-like aesthetics, suggesting enhanced commercial suitability despite minor limitations in fine-grained detail.

6 Discussion

This study systematically investigated the resource-efficient specialization of a multimodal diffusion model for minimalist logo generation. By combining LoRA-based fine-tuning with ControlNet guidance, a prototype was developed that effectively processes both textual and structural constraints. This chapter interprets the findings in the context of the research objectives outlined in Chapter 1 and discusses limitations alongside future research directions.

6.1 Interpretation of Results

The experimental results demonstrate the efficacy of the proposed hybrid approach in achieving professional-grade design automation with limited resources.

Efficacy of the Hybrid Strategy The combination of ControlNet and LoRA fine-tuning proved highly effective. ControlNet ensured structural fidelity (+32.4% SSIM), while fine-tuning successfully adapted the model to the minimalist domain, significantly improving image quality (-30.9% FID) and semantic alignment (+3.7% CLIP). Crucially, these gains were achieved with a compact dataset (1,500 images) and limited training iterations. This efficiency highlights the robustness of the approach and suggests significant untapped potential: scaling up data volume and training duration could yield even greater performance improvements without requiring industrial-grade resources.

Hyperparameter Dynamics The systematic analysis of LoRA hyperparameters revealed that the learning rate is the dominant factor for convergence and quality, with $lr = 1e-4$ consistently outperforming lower rates. High LoRA ranks ($r = 32$) combined with the "extended" module configuration (adapting both attention and feed-forward layers) provided the optimal balance, allowing the model sufficient capacity to learn domain-specific features without overfitting.

Qualitative Assessment and Practical Viability The qualitative analysis of representative case studies confirms that the fine-tuned model produces results that are stylistically superior to the base model, particularly in generating solid backgrounds and consistent, minimalist color palettes. The successful interpretation of a hand-drawn sketch (Case E4) highlights the model's practical potential for bridging the gap between rough ideation and polished design, effectively handling authentic human inputs.

6.2 Limitations and Future Work

While the results demonstrate the viability of the proposed pipeline, several limitations identify key areas for future optimization.

Data Quality and Augmentation

The dataset, while extensive, is limited to specific minimalist styles. The existing captions are often simplistic tags, lacking descriptive depth.

- **Future Work:** Future iterations should employ Vision-Language Models (VLMs) to generate ultra-detailed captions and structured categorizations (e.g., wordmarks vs. pictorial marks). This would significantly enhance the model's semantic understanding [32]. Additionally, synthetic sketch generation could be diversified using edge detection or dedicated sketch models to improve robustness via data augmentation.

Modeling Efficiency and Stability

The experiments revealed training instability characteristic of LoRA fine-tuning [31]. Hardware constraints (consumer GPU) limited batch sizes to 8.

- **Future Work:** Implementation of QLoRA (4-bit quantization) [25] and "torch.compile()" would allow for larger batch sizes and faster training on consumer hardware. Furthermore, Bayesian optimization could replace grid search to more efficiently explore the hyperparameter space and identify stable convergence regions.

Evaluation Metrics

Standard metrics (CLIP, FID, SSIM) do not fully capture design-specific criteria like memorability or vectorizability.

- **Future Work:** Developing domain-specific metrics (e.g., automated vectorization success rates) and conducting larger-scale studies with professional designers would provide more robust quality assessments.

7 Conclusion and Outlook

7.1 Summary

This thesis investigated the resource-efficient optimization of a multimodal diffusion model for minimalist logo generation. By combining LoRA-based fine-tuning with ControlNet guidance, we developed a prototype capable of operating on consumer-grade hardware (NVIDIA RTX 5080). The experimental evaluation confirmed that structural guidance is indispensable for geometric precision (SSIM +32.4%), while domain-specific fine-tuning significantly enhances semantic and visual quality (FID -30.9%). Systematic analysis identified the learning rate as the critical lever for convergence, with higher rates ($1e-4$) consistently yielding superior results.

7.2 Scientific Contributions and Implications

A key contribution of this work is the validation of parameter-efficient training strategies for the specific domain of logo design.

- **Hyperparameter Dynamics:** Our findings confirm literature recommendations for high learning rates ($1e-4$) in PEFT methods [17]. However, contrary to the suggestion that very low ranks (4 or 8) constitute a "sweet spot" [17], our experiments demonstrated that a higher rank of 32 provided the necessary capacity to capture the stylistic nuances of minimalist design without overfitting.
- **Efficiency Verification:** We demonstrated that a compact dataset of 1,500 images and approximately 2,500 training steps are sufficient to achieve professional-grade results. This challenges the assumption that massive datasets are a prerequisite for effective domain adaptation.

7.3 Future Directions

Future optimization should prioritize data quality over quantity. The integration of Vision-Language Models (VLMs) offers a promising avenue to replace simplistic tag-based captions with dense, descriptive natural language, significantly improving semantic alignment. Furthermore, advances in synthetic data generation and evolving model architectures suggest that the barrier to entry for training specialized, high-quality generative models will continue to decrease. Generative AI establishes itself not as a replacement for human expertise, but as a powerful instrument for exploration and iteration, democratizing access to professional design tools.

References

[1] R. A. Bertão, M.-H. Yeoun, and J. Joo, "A blind spot in ai-powered logo makers: Visual design principles," *Visual Communication*, vol. 24, no. 1, 2023. [Online]. Available: https://doi.org/10.1177/14703572231155593.

[2] Hugging Face, runwayml, and StabilityAI. "Stable Diffusion v1-5 Model Card." Latent text-to-image diffusion model. (2022), [Online]. Available: https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5 (visited on 2025-10-06).

[3] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *CoRR*, vol. abs/2302.05543, 2023. arXiv: 2302.05543. [Online]. Available: https://arxiv.org/abs/2302.05543.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *CoRR*, vol. abs/2204.06125, 2022. arXiv: 2204.06125. [Online]. Available: https://arxiv.org/abs/2204.06125.

[5] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *CoRR*, vol. abs/2112.10741, 2021. arXiv: 2112.10741. [Online]. Available: https://arxiv.org/abs/2112.10741.

[6] W. Lidwell, K. Holden, and J. Butler, *Universal Principles of Design*, Revised and Updated. Rockport Publishers, 2010.

[7] A. Wheeler, *Designing Brand Identity: An Essential Guide for the Whole Branding Team*, 3th. John Wiley & Sons, 2009.

[8] M. Hjalmarsson and W. Skoglund, "The impact of digitalization on logo design," M.S. thesis, Jönköping University, Jönköping International Business School, 2021. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1573306/FULLTEXT01.pdf.

[9] E. Lupton, *Thinking with Type: A Critical Guide for Designers, Writers, Editors, & Students*, 2nd. Princeton Architectural Press, 2010.

[10] RASH GRAPHIC. "7 types of logo." Artikel auf Medium über die verschiedenen Arten von Logo-Designs, inkl. Emblem Logo. (2024), [Online]. Available: https://medium.com/@rashgraphicbd/7-types-of-logo-c69496b7735d (visited on 2025-09-26).

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *CoRR*, vol. abs/1406.2661, 2014. arXiv: 1406.2661. [Online]. Available: http://arxiv.org/abs/1406.2661.

[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017. arXiv: 1701.07875.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020. arXiv: 2006.11239. [Online]. Available: https://arxiv.org/abs/2006.11239.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2022. arXiv: 2112.10752. [Online]. Available: https://arxiv.org/abs/2112.10752.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021. arXiv: 2103.00020.

[16] D. Zhang, T. Feng, L. Xue, Y. Wang, Y. Dong, and J. Tang, "Parameter-efficient fine-tuning for foundation models," 2025. arXiv: 2501.13787v1. [Online]. Available: https://Awesome-PEFT-for-Foundation-Models.github.io.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Han, W. Chen, and L. Wang, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. arXiv: 2106.09685. [Online]. Available: https://arxiv.org/abs/2106.09685.

[18] C. Zhihao, S. Zhihao, M. Li, W. Jiahao, Z. Yifan, L. Xinyi, C. Wei, and S. Qian, "A comprehensive survey of parameter-efficient fine-tuning for large language and vision models," 2025. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.175303890.08902658/v1.

[19] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *ArXiv*, vol. abs/2104.08718, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233296711.

[20] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, pp. 600–612, 2004.

[21] J. Snell, K. Swersky, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1112–1121.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] iamkaikai. "amazing_logos_v4 dataset." version v4. (2024), [Online]. Available: https://huggingface.co/datasets/iamkaikai/amazing_logos_v4 (visited on 2025-09-01).

[24] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, *Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation*, 2023. arXiv: 2208.12242 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2208.12242.

[25] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *Qlora: Efficient finetuning of quantized llms*, 2023. arXiv: 2305.14314 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2305.14314.

[26] MLflow Team. "Mlflow." (2024), [Online]. Available: https://mlflow.org/ (visited on 2025-10-10).

[27] Cloneofsimo (GitHub Author). "Lora: Low-rank adaptation for fast text-to-image diffusion fine-tuning." (2023), [Online]. Available: https://github.com/cloneofsimo/lora (visited on 2025-10-22).

[28] OpenAI. "Clip: Connecting text and images." (2021), [Online]. Available: https://github.com/openai/CLIP (visited on 2025-10-10).

[29] scikit-image Team. "Scikit-image." (2024), [Online]. Available: https://scikit-image.org/ (visited on 2025-10-10).

[30] G. Parrish. "Clean-fid." (2021), [Online]. Available: https://github.com/GaParmar/clean-fid (visited on 2025-10-10).

[31] Z. Luo, X. Xu, F. Liu, Y. S. Koh, D. Wang, and J. Zhang, *Privacy-preserving low-rank adaptation against membership inference attacks for latent diffusion models*, 2024. arXiv: 2402.11989 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2402.11989.

[32] Y. Zeng, Y. Qi, Y. Zhao, X. Bao, L. Chen, Z. Chen, S. Huang, J. Zhao, and F. Zhao, "Enhancing large vision-language models with ultra-detailed image caption generation," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds., Suzhou, China: Association for Computational Linguistics, 2025, pp. 26 703–26 729, ISBN: 979-8-89176-332-6. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1357/.