A Multi-Stage Approach to Image Consistency in Zero-Shot Character Art Generation for the D&D Domain

Gayashan Weerasundara & Nisansa de Silva

Content

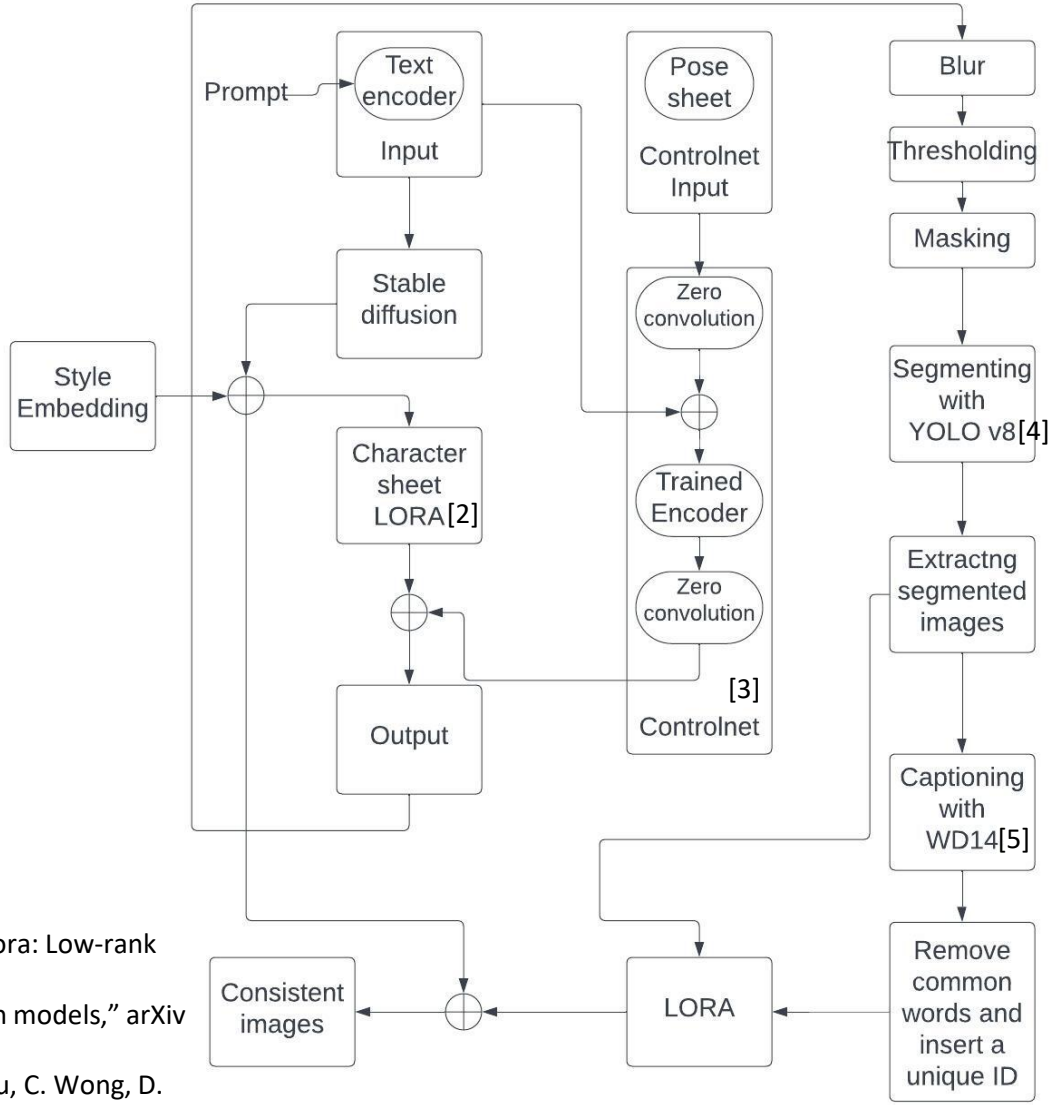
- Introduction
- Overall Pipeline
- Training a General Art Style
- Training a Specific Character Generation
- Character Sheet Generation
- Separation and Training for Specific Character
- Workflow
- Results
- Specifications
- Q&A

Introduction

- In the evolving landscape of computer graphics, the pursuit of consistency in image generation remains a compelling challenge. This paper delineates a comprehensive methodology that harnesses the capabilities of the Stable Diffusion (SD) model [1], an adept open-source image generation tool, to generate consistent and high-fidelity imagery in a zero shot manner.

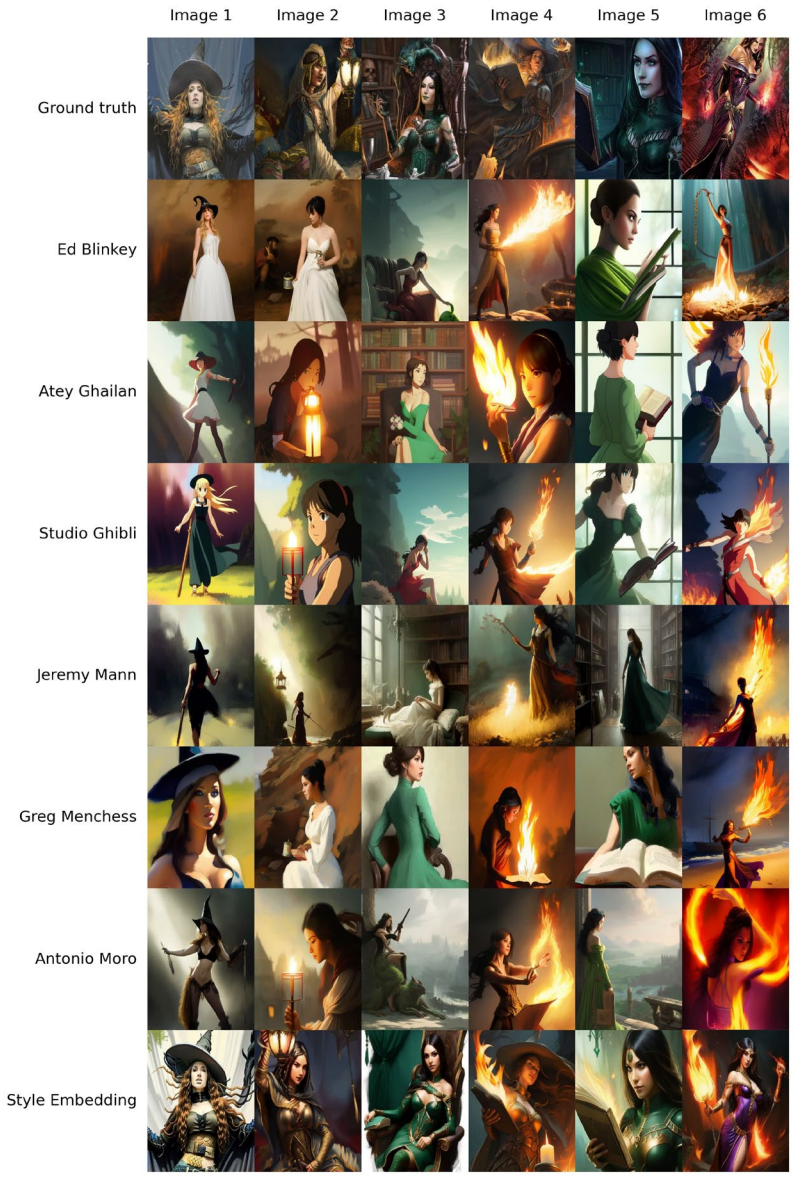
1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Proposed Pipeline

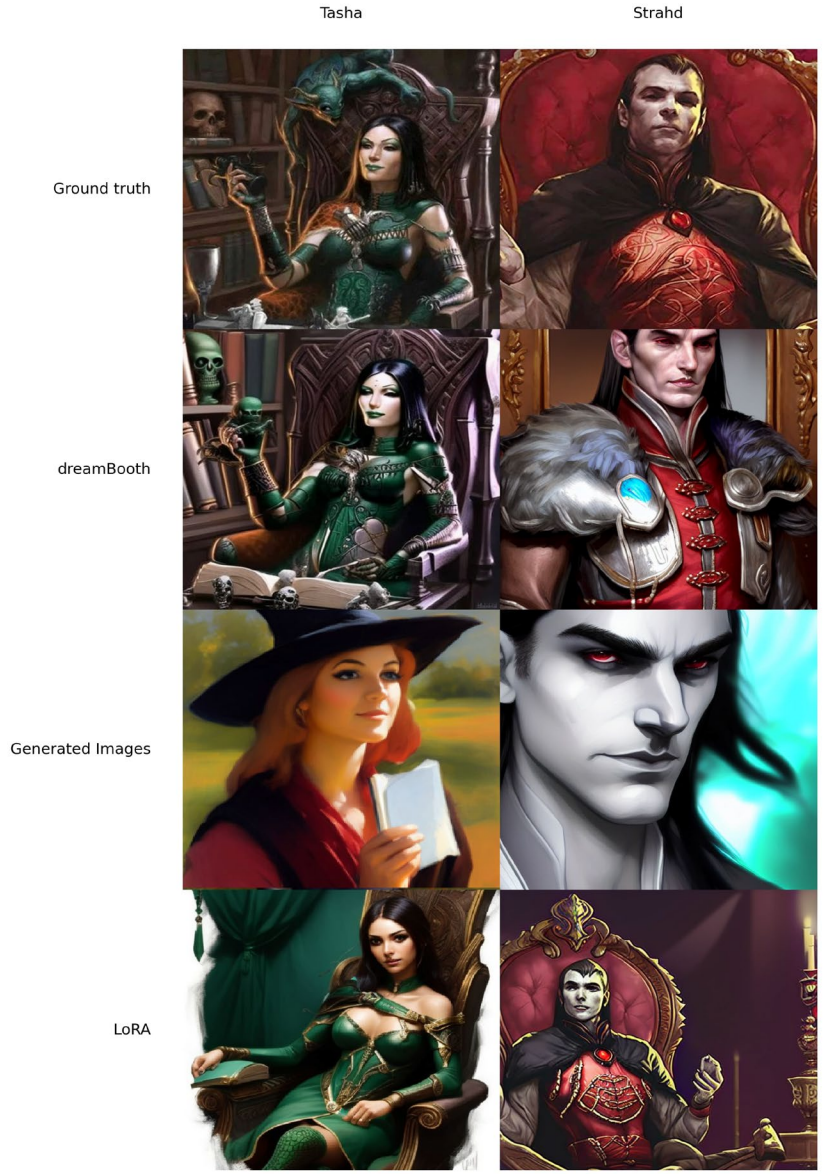


2. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” arXiv preprint arXiv:2106.09685, 2021.
3. L. Zhang and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” arXiv preprint arXiv:2302.05543, 2023.
4. G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, Z. Yifu, C. Wong, D. Montes et al., “ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation,” Zenodo, 2022.
5. https://bit.ly/3RuKhc7

Training a General Art Style

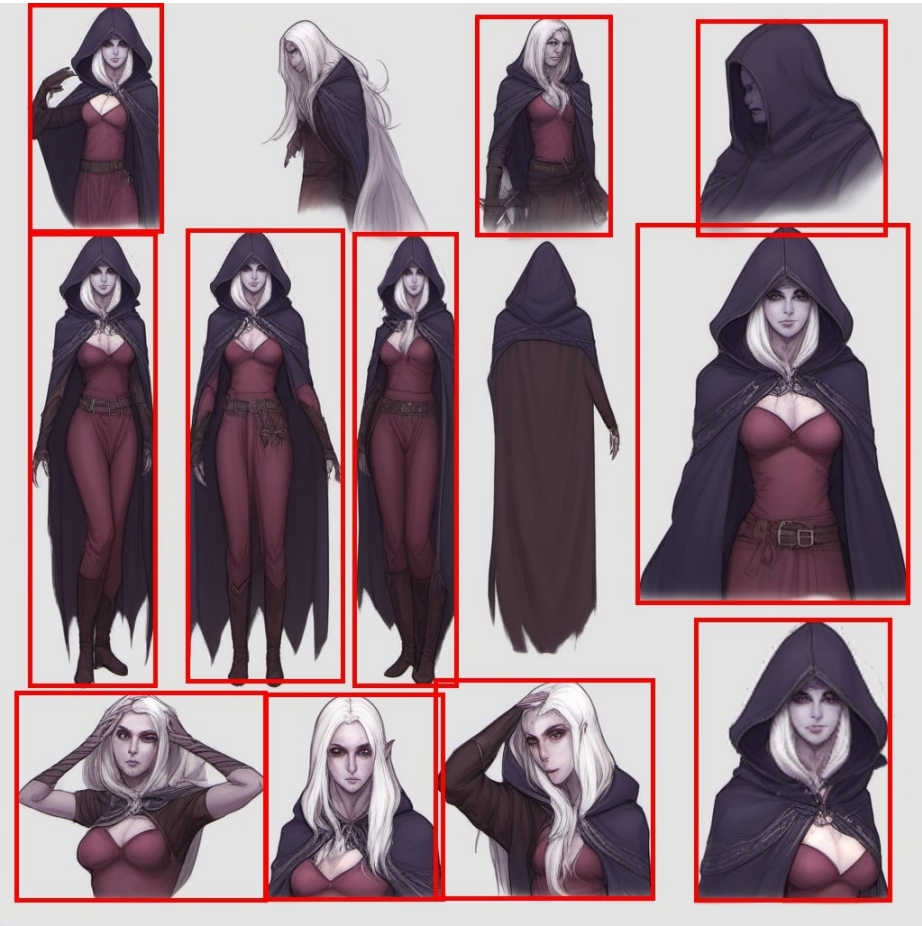


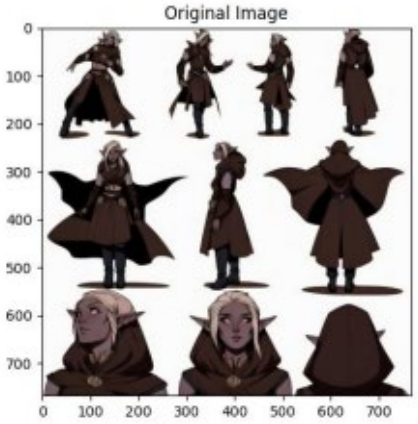
Training a Specific Character Generation

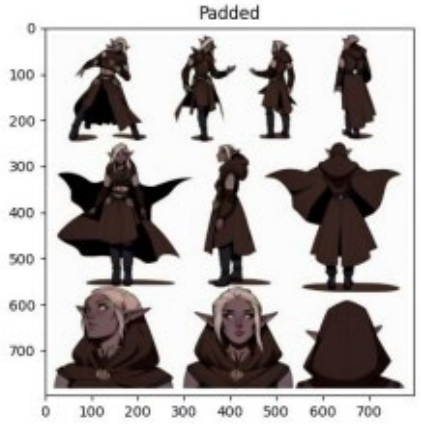


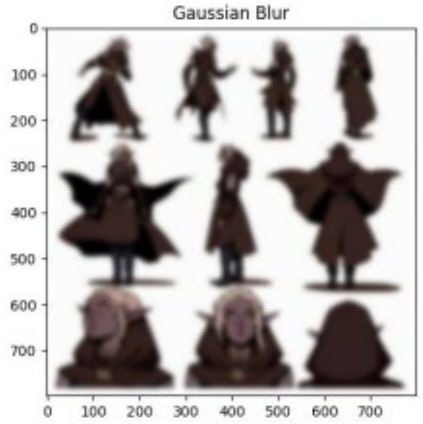
Character Sheet Generation

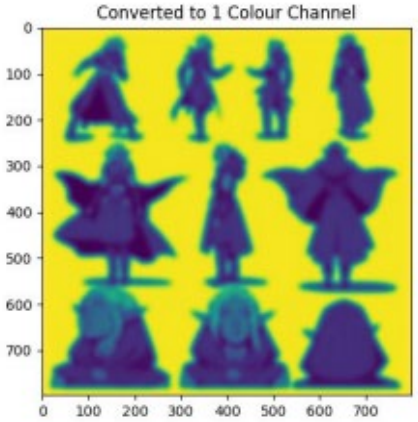
Separation and Training for Specific Character

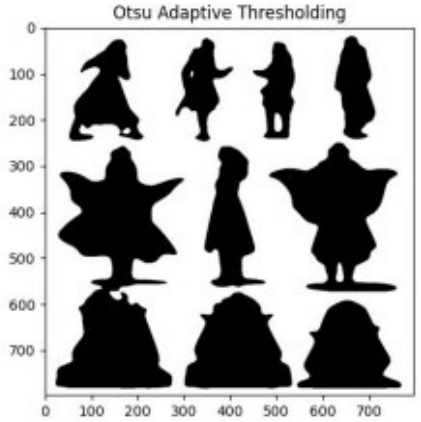


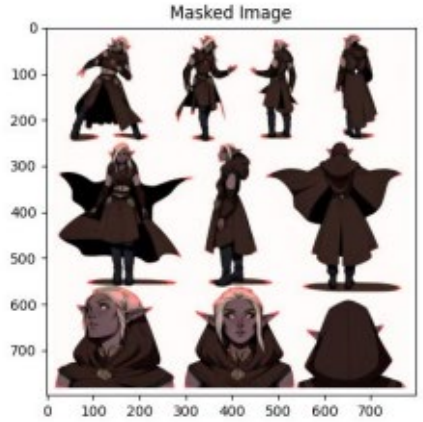












Workflow

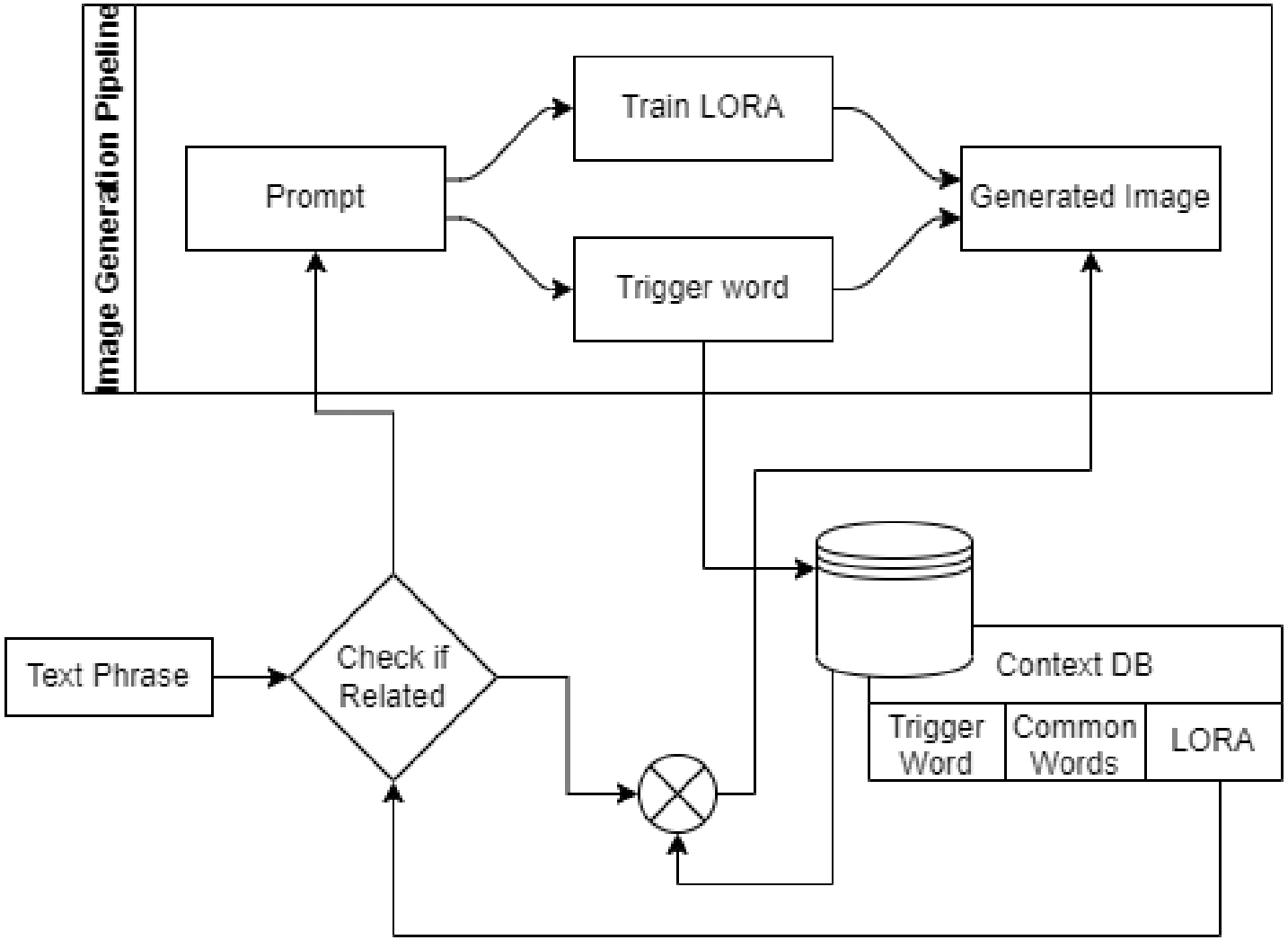


Image Generation Pipeline
Prompt
Train LORA
Trigger word
Generated Image
Text Phrase
Check if Related
Context DB
Trigger Word
Common Words
LORA

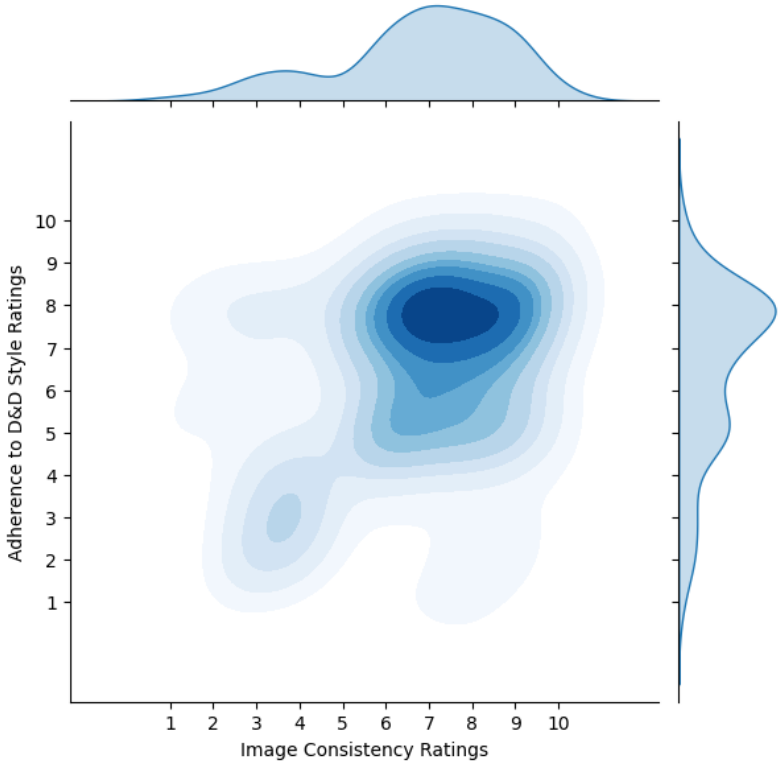
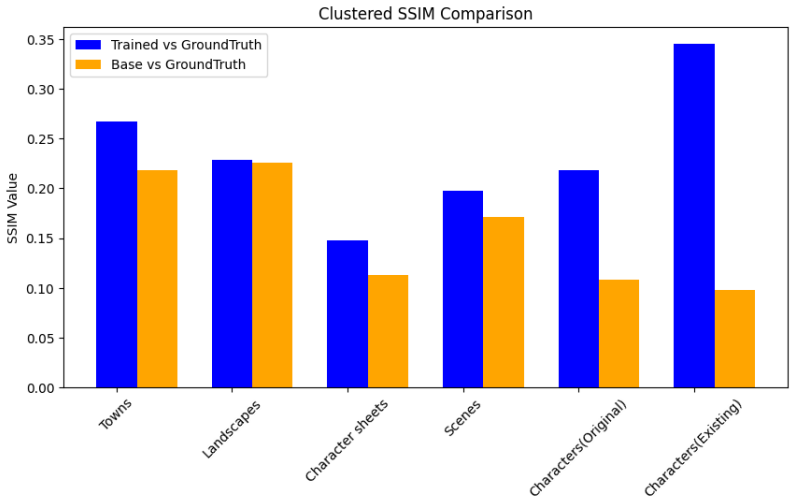


	Image to Image	InstructPix2Pix	Proposed Method
LLaVA (Liu et al., 2023)	6.4	7.5	8.3
FID (Heusel et al., 2017)	51.7	38.9	13.2







Results

6. H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” arXiv preprint arXiv:2304.08485, 2023
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

Specifications

Parameter	Pre-trained Character LoRA	Zero-Shot Character LoRA	D&D Style LoRA	Character-sheet LoRA
Number of Images	40 - 60	4 - 12	676	488
Repeats	10 - 15	20 - 40	2	4
Epochs	15 - 25	20 - 30	12	15
Precision	bf16	bf16	bf16	bf16
Learning Rate	0.0001	0.0001	0.0003	0.0002
Warmup	0	0	0 - 5%	0 - 5%
Resolution	1024x1024	1024x1024	1024x1024	1024x1024
Optimizer	Adafactor	Adafactor	Adafactor	Adafactor
Batch Size	2 - 4	1	4	4
Gradient Checkpointing	True	True	True	True
Xformers	False	False	False	False

Table 1: Training configurations for LoRA models trained for the proposed pipeline

Component	Specification
GPU	Nvidia RTX 4090
VRAM	24GB
RAM	61GB
CPU	16vCPU

References

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
2. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” arXiv preprint arXiv:2106.09685, 2021.
3. L. Zhang and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” arXiv preprint arXiv:2302.05543, 2023.
4. G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, Z. Yifu, C. Wong, D. Montes et al., “ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation,” Zenodo, 2022.
5. https://bit.ly/3RuKhc7
6. H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” arXiv preprint arXiv:2304.08485, 2023
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.