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

Gayashan Weerasundara[a] and Nisansa de Silva[b]
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

Keywords: Machine Learning, Image Generation, Consistency, Dungeons & Dragons.

Abstract: 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, an adept open-source image generation tool, to generate consistent and high-fidelity imagery. Beginning with the training of a generalized art style for the popular tabletop role-playing game Dungeons and Dragons, our research journeyed through specific character generation and the creation of exhaustive character sheets, culminating in the generation of consistent character images across various poses.

1 INTRODUCTION

The computer graphics field has recently transformed with new image generation techniques, emphasizing realism and consistency, crucial in gaming, animation, and virtual reality. This is particularly relevant for the Dungeons & Dragons (D&D) universe, where consistent, high-fidelity character images are vital to maintaining its authentic high-fantasy (Peiris and de Silva, 2022) atmosphere. However, most recent studies (Peiris and de Silva, 2023; Weerasundara and de Silva, 2023) focused on natural language processing aspects of the game rather than image generation. Recent strides in diffusion models, especially the Stable Diffusion (SD) model (Rombach et al., 2022), noted for producing quality images consistently (Ho et al., 2020), have not been fully explored in D&D character generation.

Our paper explores using the SD model to create consistent, high-quality D&D character images. We discuss training a generalized art style, transitioning to character generation, and producing character sheets and consistent poses. Our approach combines recent research with the SD model's capabilities, aiming to advance computer graphics applications, particularly in the D&D realm. We detail our methodology, compare it with existing techniques, and analyze our results. Our goal is to present a new approach to computer graphics, focusing on the D&D universe.

2 RELATED WORK

In the evolving field of text-to-image synthesis and story visualization, several key studies have made notable contributions. *StoryDALL-E* by Maharana et al. (2022) is pivotal, focusing on enhancing large transformers for story visualization. They introduced story continuation and evaluated it on datasets such as PororoSV (Li et al., 2019), FlintstonesSV (Gupta et al., 2018), and DiDeMoSV (Anne Hendricks et al., 2017), achieving success in sequential image generation.

Jeong et al. (2023) presents a novel method for creating storybooks from text using diffusion models. Rahman et al. (2023) contributes to generating visually consistent stories based on visual memory. Pan et al. (2022) explores coherent story creation using latent diffusion models. Cho et al. (2023) work delves into the reasoning and biases of generative transformers. The work by Esser et al. (2021) focuses on high-resolution image synthesis using transformers, while Liang et al. (2019) discuss content parsing in text-to-image synthesis through *CPGAN*.

Other studies have also advanced the field, including Maharana et al. (2021)'s work on semantic consistency in visual stories, Frans et al. (2022)'s exploration of text-to-drawing synthesis, and Wang et al. (2018)'s development of high-resolution image synthesis models. These studies collectively form the foundation of our research, influencing our approach and objectives in text-to-image synthesis and story visualization.

[a] https://orcid.org/0000-0003-1419-8938
[b] https://orcid.org/0000-0002-5361-4810

Weerasundara, G. and de Silva, N.
A Multi-Stage Approach to Image Consistency in Zero-Shot Character Art Generation for the D&D Domain.
DOI: 10.5220/0012317600003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 235-242
ISBN: 978-989-758-680-4; ISSN: 2184-433X

3 METHODOLOGY

The process of generating consistent characters can be broken down to five main steps.

3.1 Training a General Art Style

Our research began by embedding a generalized art style into the diffusion model, crucial for setting the aesthetic for later stages. We collected a diverse dataset through web scraping and from official D&D books, focusing on the unique art style of D&D—a mix of realism and fantasy.

We curated the dataset to ensure each image matched our aesthetic goal, which resembled official D&D art, capturing the essence of oil and watercolor paintings. This curation was vital for training our model to create an aesthetic embedding for style transfer, central to our research.

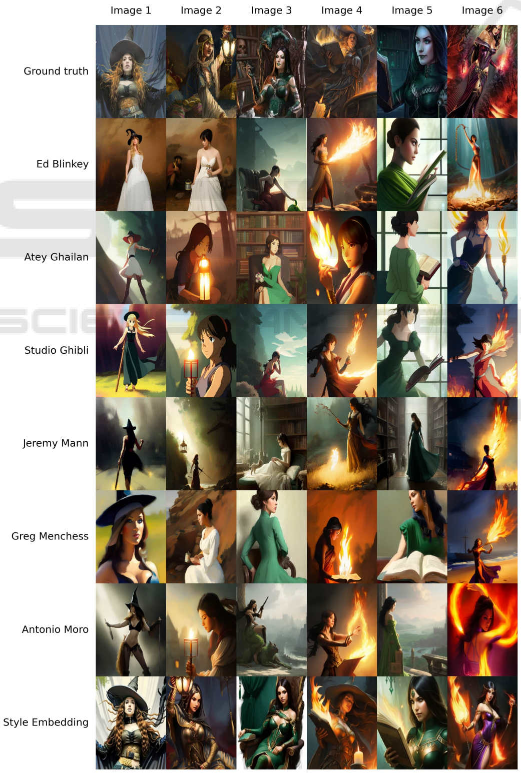

Figure 1: Comparing official images, style embedding, and generated images using various artists as basis.

Figure 1 demonstrates the effectiveness of our training. It compares official and fan-made D&D images with those generated by our model, using various artist styles. Our approach avoided negative prompts, allowing for a natural representation of the chosen style. The results show our method's efficiency and the potential of merging artistic elements with computational techniques.

3.2 Training a Specific Character Generation

Our work honed in on specific character image generation within the domain of the Stable Diffusion (SD) model, utilizing two advanced techniques: *DreamBooth* (Ruiz et al., 2023) and *LoRA* (Low-Rank Adaptation) (Hu et al., 2021), each known for their image generation capabilities. We crafted a curated dataset from web-scraped images and free official D&D sources such as dndBeyond, focusing on two D&D characters: *Tasha* from *Tasha's Cauldron of Everything* (Crawford et al., 2020), and *Strahd* from *Curse of Strahd* (Perkins et al., 2016).

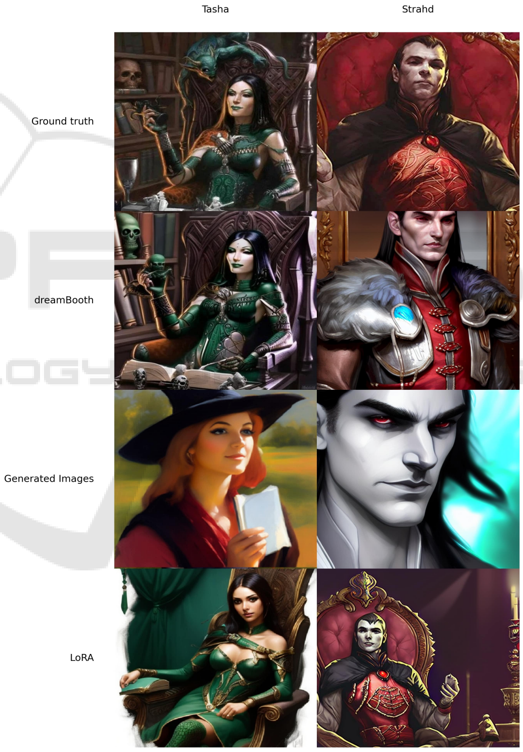

Figure 2: Comparing official character art, and images generated from *LoRA*, *DreamBooth*, and the base model.

The *DreamBooth* technique was applied directly to the images for training. LoRA's approach, in contrast, required detailed descriptions for each image, omitting the character and using a keyword instead. For this, we used the *BLIP* (Li et al., 2022) image captioning system to create preliminary captions, which were then carefully refined to align with the images.

Figure 2 offers a visual representation of our en-

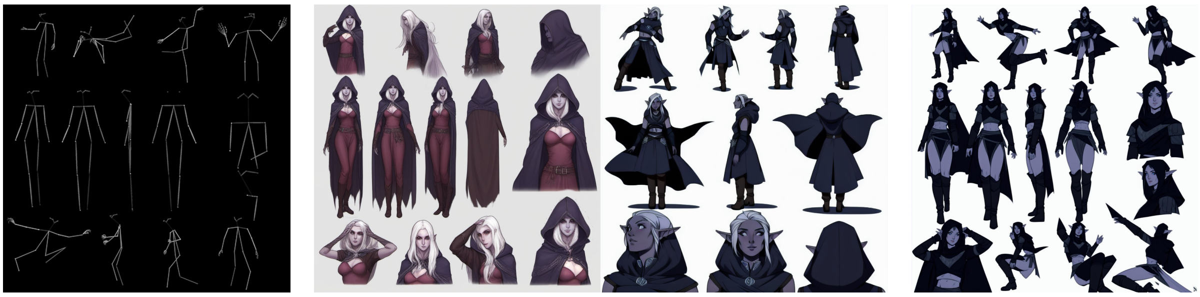

Figure 3: Character sheets generated by a given prompt using *ControlNet*.

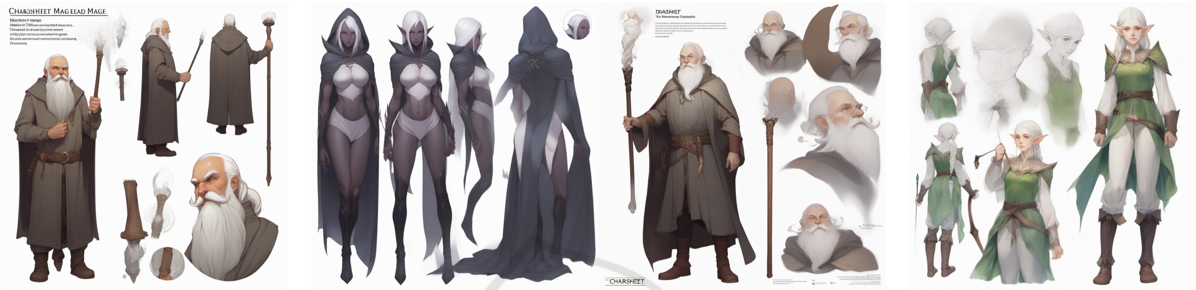

Figure 4: Character sheets generated using the trained *LORA*.

deavours. It juxtaposes the images generated using *LoRA*, *DreamBooth*, and the base model against the official character art of *Tasha* and *Strahd*. The results were illuminating. The *LoRA* model, in particular, stood out for its exceptional capability.

3.3 Character Sheet Generation

Leveraging a generated character image, our objective was to create a comprehensive character sheet using a specialized *LORA* and a *ControlNet* (Zhang and Agrawala, 2023).

3.3.1 ControlNet

We selected `controlnet-openpose-sdxl-1.0`[1] due to its bias towards generating full-body images. First, we generated an image of a character and then keeping the seed value, prompt, scale and negative embedding consistent, we provided a reference pose sheet to the *ControlNet*. The system then extrapolated this image, generating a detailed character sheet inclusive of different poses and variations.

Figure 3 shows character sheets generated using *ControlNet*. Generated images consist of specific characters in various poses in a blank background. The first image shows a sample reference sheet.

3.3.2 Training a LORA

While *ControlNet* allows the generation of character sheets, it is not consistent. In a random generation, it is possible to get distorted images very easily. To make the generation more consistent, we need to embed the concept of a character sheet into the generated images. For that, we used a *LORA* model. To train a *LORA*, we need to create a labelled dataset of reference images. We used web scraping on Google images with keywords such as *Character sheet* and *reference sheet* and manually filtered undesired images. Resultant images were captioned with *WD14*[2] tagger to get preliminary labels. Next tags which are common to most of the images such as *white background*, *Character sheet*, *reference sheet*, *turnover* are removed while adding a common unique identifier as a trigger word. Finally, a *LORA* for that specific character can be trained on the labelled image dataset. Results for character sheet *LORA* can be seen in Figure 4

3.4 Separation and Training for Specific Character

Upon obtaining the character sheet, our focus shifted towards extracting undistorted and apt images that showcase the character in a variety of poses. This ex-

[1] https://bit.ly/3RKSPNa

[2] https://bit.ly/3RuKhc7

traction process is crucial as it provides the necessary training data for developing a dedicated model capable of generating images of the character in specific, predefined poses.

3.4.1 Image Processing for Segmentation

The initial step in this process involved a series of image-processing techniques aimed at isolating each image from the character sheet. The image was first padded to ensure that the boundaries of each character were well-defined. To reduce artifacting in the subsequent steps, Gaussian blur was applied to the image. Following this, the image was converted to a single colour channel, and Otsu thresholding was applied to create a binary mask. This mask was then overlaid on top of the padded image with a low alpha value, creating a masking effect that highlights each character distinctly. Figure 5 illustrates the sequence of image processing techniques employed to segment the images on the character sheet.

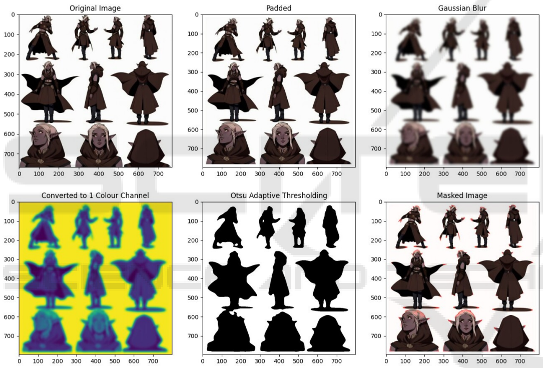

Figure 5: Sequence of image processing techniques employed for segmenting images.

3.4.2 Object Detection for Image Separation

Following the image processing, a fine-tuned `YOLO v8` (Jocher et al., 2022) object detection model was employed as a segmentation model to separate the masked images. The model was adept at identifying and isolating the individual characters highlighted by the mask, where the regions of interest were then mapped against the original images to extract the character representations in various poses.

3.4.3 Model Training and Image Generation

The segmented images served as the training data for a dedicated model, designed to generate images of the character in various predefined poses. The training process was meticulous, ensuring that the model learned the nuances of the character's appearance and pose variations. The trained model demonstrated proficiency in generating consistent, high-quality images that accurately represented the character in the specified poses.

3.5 Generating Consistent Characters Using the Trained Model

The culmination of our research efforts was the generation of consistent character images using the trained model. We used a combination of trained LORA models alongside style embedding trained on DnD images to archive consistent character generation.

The primary objective was to ensure that the generated images maintained a high degree of consistency in terms of character details, pose accuracy, and overall visual quality. Using the trained model, we tested multiple prompts under different seed values. The seed values were varied to introduce randomness in the generation process, ensuring that the model's robustness and consistency were thoroughly tested. Each prompt was designed to elicit a specific response from the model, be it a particular pose, expression, or background setting.

The generated images showcased remarkable consistency as shown in Figure 6. The characters were homogeneous in style, and the poses were accurate, confirming the efficacy of our training process. Furthermore, the visual quality of the images was high, with sharp details and vibrant colours, making them suitable for professional use in storyboards, animations, and other media.

3.6 Saving Context for Re-Usability

A pivotal aspect of our methodology is the incorporation of a mechanism for saving context, which significantly enhances the re-usability of the trained LORA models. This mechanism is designed to capture and store essential contextual information during the training of a LORA model, thereby facilitating its subsequent utilization in generating images that align with specific character descriptions.

3.6.1 Context Capturing and Storage

During the training of a LORA model, images are captioned using the *WD14* tagger. This process results in the generation of a set of tags that describe the common features across every image, such as *white hair*, *old man*, *white beard*, which are indicative of the character's attributes. These common tags are then replaced with a single unique trigger word, serving as an identifier for the character's context.

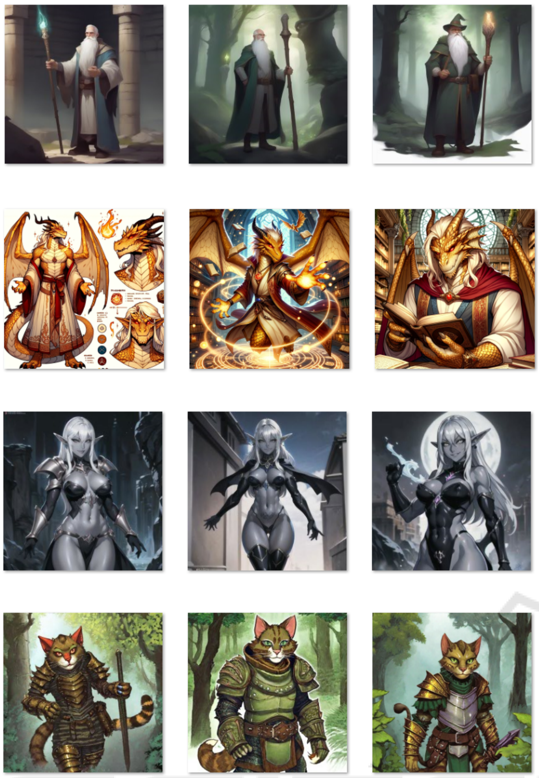

Figure 6: Examples showcasing the consistency in generated characters.

The captured context, comprising the trigger word, the common words (tags), and the trained *LORA* `SafeTensor`, is systematically stored in a dedicated database, referred to as the *context database* (`context db`). This structured storage enables the efficient retrieval and re-use of context for generating consistent character images.

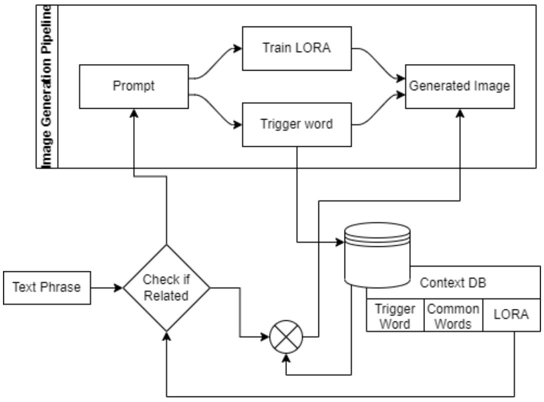

Figure 7: Workflow illustrating the process of context re-usability.

3.6.2 Context Re-Usability Workflow

Figure 7 depicts the workflow for context re-usability. Upon receiving text input, the system evaluates whether the text contains similarities to the common words stored in the context db. As per our previous study, which was conducted on extracting named entities for D&D domain Weerasundara and de Silva (2023), it is possible to identify whether a named character is included in the prompt and match them with the context db. If a match is found, the corresponding LORA model, identified by the trigger word, is utilized to generate images that adhere to the character's context. This approach eliminates the need for re-training a LORA model for known contexts, thereby optimizing the image generation process.

In scenarios where the text input does not match any stored context, the system follows the established pipeline to train a new LORA model. The newly acquired context is then added to the context db, expanding the repository of reusable contexts for future image generation tasks.

3.6.3 Implications and Efficiency

The implementation of context saving and re-usability significantly contributes to the efficiency and versatility of our image generation system. By leveraging stored contexts, the system can rapidly generate images that are contextually consistent with known character descriptions, thereby reducing computational overhead and response time. Furthermore, the continuous expansion of the context db ensures the adaptability of the system to diverse character contexts, making it a valuable asset in applications such as gaming, animation, and virtual reality.

4 EXPERIMENTAL SETUP

In this study, we meticulously designed an experimental framework to validate the efficacy of our proposed methodology. Figure 8 illustrates the comprehensive pipeline of our approach, encompassing the initial character generation, training of character-specific LORA, and the final generation of consistent images.

Our experimental setup was configured to ensure the reproducibility and reliability of the results. The models were trained on RTX 4090 with 24GB VRAM, 61GB RAM, 16vCPU on the cloud service

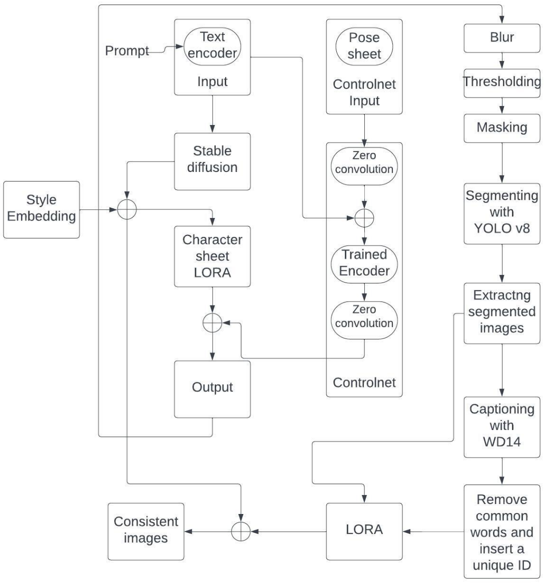

Figure 8: Architecture of the pipeline used for the system.

runpod.io[3] using the `Kohya ss GUI`[4]. The parameters were fine-tuned, and multiple iterations were conducted to optimize the performance of the Stable Diffusion (SD) model and the LORA model.

5 RESULTS

The results of our experiments were promising, showcasing the capability of our methodology in generating high-fidelity and consistent character images. The generated images exhibited remarkable consistency in style, pose accuracy, and visual quality, adhering to the predefined artistic style of the D&D universe.

To validate the effectiveness of our approach, we conducted a comparative study against established benchmarks, focusing on the structural similarity index measure (SSIM) (Zermani et al., 2021). SSIM is a widely recognized metric for comparing the similarity between two images, providing insights into the perceptual changes between the generated images and the ground truth.

Figure 9 presents the average SSIM comparison between the official images, baseline generations, and the images generated using the trained model per each category of images which indicates a greater consistency between official art with trained models compared to base models.

In our study, we evaluated the consistency and D&D style adherence of generated images through an online survey with 291 participants. They rated the consistency of characters across multiple images and the stylistic alignment with traditional D&D imagery on a scale of 1 to 10.

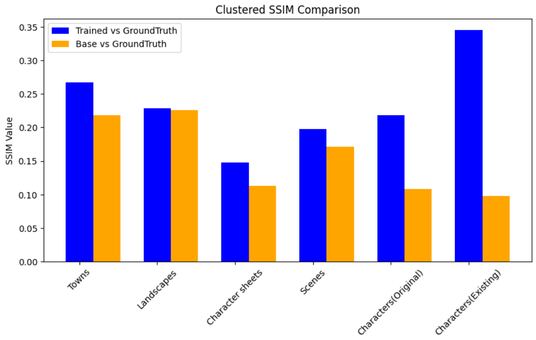

Figure 9: SSIM Comparison between ground truth, and images generated from trained LoRA, and the base model.

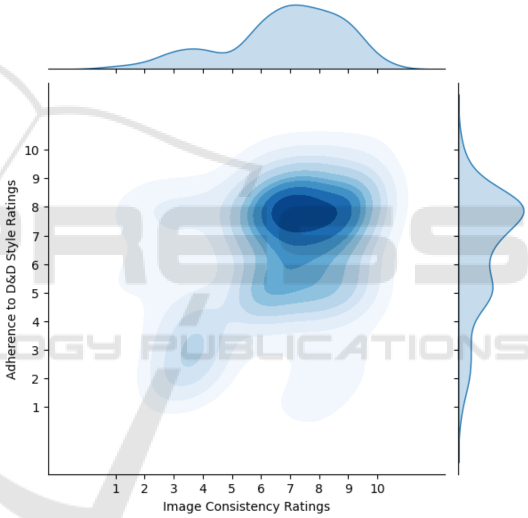

Figure 10: Density Distribution of Feedback on Generated Images.

Figure 10 density plot synthesizes these ratings. The x-axis shows character consistency ratings and the y-axis indicates style adherence. The darker areas represent a higher consensus among participants. The results demonstrate that our method effectively produces images that are both consistent in character depiction and faithful to the D&D style, as indicated by the concentration of responses in the upper regions of the plot.

The plot reveals a notable clustering of responses in the high-density regions, suggesting a collective agreement among participants. The results indicate a significant trend wherein our image generation approach has successfully maintained a high level of consistency in character portrayal and adherence to

[3] https://www.runpod.io/

[4] https://github.com/bmaltais/kohya_ss

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

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

the D&D style. The darker regions of the plot signify a concentration of responses indicating both high consistency and high stylistic fidelity, thereby affirming the effectiveness of our generative model in producing images that align with the expectations of D&D imagery.

We have also conducted a comparison using the LLaVA (Liu et al., 2023) model and the FID (Heusel et al., 2017) score. LLaVA is an open-source multi-model that has both vision capabilities and LLM capabilities, thus act as an ideal agent for comparative tasks. By giving a series of images combined as a single image, the LLaVA model can be asked to provide a similarity score (1-10) while asking specifically to provide only the score. Averages of resultant scores are mentioned in Table 2.

Table 2: Comparison between stable diffusion image to image conversion, *InstructPix2Pix* (Brooks et al., 2022) method and the proposed method.

	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

From Table 2, it can be seen that the performance of the proposed methodology in generating consistent images is remarkably high.

6 DISCUSSION

Our research focused on leveraging the Stable Diffusion (SD) model to generate consistent and high-fidelity images of characters within the Dungeons & Dragons universe. We successfully utilized advanced techniques like *LoRA* and *ControlNet* to create visually appealing and consistent character images across different poses and scenarios, primarily targeting applications in gaming, animation, and virtual reality. Comparative analysis using benchmarks such as *SSIM* and human validation indicated that our approach maintained structural similarity with ground truth and outperformed baseline models, demonstrating its potential in the field of computer graphics.

However, our research faced challenges in curating a diverse dataset for style embedding and character sheet *LoRA*. Additionally, the generation of character sheets sometimes resulted in distortions due to the high number of variations. Despite these challenges, our findings hold significant promise for future developments in the image generation field.

7 CONCLUSION

In conclusion, our research delineated a comprehensive approach for generating consistent and high-fidelity character images using the Stable Diffusion model. The results demonstrated the potential of our methodology to contribute to the evolving landscape of computer graphics. While challenges were encountered, the insights gained and the avenues opened for future work make this research a valuable addition to the field.

8 FUTURE WORK

The findings of our research open several avenues for future exploration. The refinement of the dataset curation process, particularly through the integration of the proposed method in Avrahami et al. (2023), we can more effectively refine the images selected for training, potentially improving both the quality and consistency of generated images. Additionally, the application of our methodology to genres beyond Dungeons & Dragons presents an exciting prospect.

REFERENCES

Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. (2017). Localizing moments in video with natural language. In *Proceedings* *of the IEEE international conference on computer vision*, pages 5803–5812.

Avrahami, O., Hertz, A., Vinker, Y., Arar, M., Fruchter, S., Fried, O., Cohen-Or, D., and Lischinski, D. (2023). The chosen one: Consistent characters in text-to-image diffusion models. *arXiv preprint arXiv:2311.10093*.

Brooks, T., Holynski, A., and Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*.

Cho, J., Zala, A., and Bansal, M. (2023). Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Crawford, J., Dillon, D., Petrisor, B., Schneider, F. W., and Teague, E. (2020). *Tasha's Cauldron of Everything*. Wizards of the Coast Publishing.

Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Frans, K., Soros, L., and Witkowski, O. (2022). Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *NeurIPS*, 35:5207–5218.

Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., and Kembhavi, A. (2018). Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jeong, H., Kwon, G., and Ye, J. C. (2023). Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Yifu, Z., Wong, C., Montes, D., et al. (2022). ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., and Gao, J. (2019). Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338.

Liang, J., Pei, W., and Lu, F. (2019). Cpgan: full-spectrum content-parsing generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:1912.08562*.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Maharana, A., Hannan, D., and Bansal, M. (2021). Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*.

Maharana, A., Hannan, D., and Bansal, M. (2022). Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pages 70–87. Springer.

Pan, X., Qin, P., Li, Y., Xue, H., and Chen, W. (2022). Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*.

Peiris, A. and de Silva, N. (2022). Synthesis and evaluation of a domain-specific large data set for dungeons & dragons. *arXiv preprint arXiv:2212.09080*.

Peiris, A. and de Silva, N. (2023). SHADE: semantic hypernym annotator for Domain-Specific entities - DnD domain use case. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, page 6, Peradeniya, Sri Lanka.

Perkins, C., Hickman, T., and Hickman, L. (2016). *Curse of Strahd*. Wizards of the Coast.

Rahman, T., Lee, H.-Y., Ren, J., Tulyakov, S., Mahajan, S., and Sigal, L. (2023). Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2493–2502.

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

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807.

Weerasundara, G. and de Silva, N. (2023). Comparative analysis of named entity recognition in the dungeons and dragons domain.

Zermani, M., Larabi, M.-C., and Fernandez-Maloigne, C. (2021). A comprehensive assessment of the structural similarity index. *Signal Processing: Image Communication*, 99:116336.

Zhang, L. and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.