Genshin-Style Character Generation Using Diffusion Model

Xiaohan Tang* **Ruili Xu†** **Lawrence Fu‡**

Abstract

We investigate whether a generative model can exhibit creativity while operating under explicit stylistic and semantic constraints. To study this, we develop a system that generates novel yet consistent visual designs for *Genshin Impact*–style characters. A Stable Diffusion model is fine-tuned using Low-Rank Adaptation (LoRA) to learn the Genshin art style, while a multilayer perceptron (MLP) maps structured character attributes such as *region*, *vision*, and *rarity* into CLIP's text embedding space. The MLP's output provides semantic conditioning that guides the LoRA-refined diffusion model during image generation. Our results aim to empirically evaluate how generative models can balance creative diversity with adherence to strict visual and contextual rules, offering insights into creativity under constraint in modern AI systems.

1 Introduction

Whether generative models can truly be *creative* remains a controversial question. Creativity is typically defined as the ability to produce novel yet meaningful outcomes within given constraints. This is an ability humans display naturally but AI struggles to emulate. As AI continues to transform creative industries such as illustration, music, and game design, this question becomes increasingly relevant. For example, companies like HoYoverse, known for the distinctive art style of *Genshin Impact*, may one day adopt AI-assisted pipelines to speed up production. This raises an important question: can AI generate creative designs that stay stylistically consistent while accurately reflecting newly defined character attributes to preserve the game's identity and meet player expectations?

To explore this, our project develops a system that, given a set of character attributes such as *region*, *vision*, and *rarity*, generates a corresponding official character portrait in the *Genshin Impact* art style. The generation process operates under two explicit constraints:

1. **Artistic constraint:** the generated image must remain consistent with the Genshin art style, characterized by clean line art, soft gradients, and fantasy-inspired visual motifs.
2. **Semantic constraint:** structured attributes such as *region*, *vision*, and *rarity* influence visual features like color palette, accessories, and symbols.

We adopt a hybrid generative approach in which a character-attributes-to-embedding MLP guides a LoRA-refined Stable Diffusion model. LoRA captures stylistic fidelity by fine-tuning the diffusion model on existing artwork, while the MLP provides semantic conditioning by embedding structured attributes into CLIP's text space. Together, they form a controlled generation framework that enables the study of creativity within well-defined boundaries.

*isabella.tang@mail.utoronto.ca
†ruili.xu@mail.utoronto.ca
‡lawrence.fu@mail.utoronto.ca

2 Background and Related Work

2.1 CLIP and Multimodal Embeddings

Contrastive Language–Image Pretraining (CLIP) [1] defines a shared embedding space that aligns visual and textual modalities. It trains an image encoder and a text encoder jointly using contrastive loss so that semantically corresponding image and text pairs have high cosine similarity. In this space, CLIP can measure similarity between image, text, and image–text pairs, capturing both semantic meaning and compositional relationships such as "dog wearing a hat" minus "dog" is approximately equal to "hat." Because CLIP embeddings encode rich semantic information, they are frequently used to condition generative models on text or other structured inputs. For this project, we will be using CLIP to encode both images and textual character attributes into a common embedding space. This step is essential for correlating character attributes with corresponding visual outputs, enabling text-guided image generation.

2.2 Stable Diffusion for Text-to-Image Generation

Stable Diffusion [2] is a latent diffusion model that generates images by denoising a random noise latent under the guidance of a text embedding. Rather than operating directly in *pixel* space, the model iteratively refines a *latent* representation, which is then decoded into a final image using a variational autoencoder (VAE). During inference, the text prompt is converted into an embedding through a pretrained text encoder, often CLIP's, and this embedding steers the denoising process to produce images that are semantically aligned with the prompt. This architecture forms the foundation of many modern text-to-image systems because it is both efficient and capable of high-quality results.

2.3 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) [3] is an efficient method for fine-tuning large diffusion models. Instead of retraining the full set of model weights, LoRA inserts trainable low-rank matrices into attention and projection layers, which significantly reduces the number of parameters that must be updated. For this project, we will be using a diffusion model with LoRA fine-tuning. This approach ties directly into our resource constraints, because LoRA freezes most pretrained weights and allows us to adapt a large model efficiently with limited time and computational resources. When trained on paired image and caption data, LoRA allows Stable Diffusion to learn a specific aesthetic style, such as Genshin character designs, without requiring extensive compute or data.

2.4 Related Work

While no previous research has specifically explored image generation for *Genshin Impact* characters, the broader problem of style-aligned image generation has been studied. The authors of Style-Aligned Image Generation via Shared Attention [4] proposed a diffusion-based approach that uses a style reference image to guide the denoising process through shared attention layers. Although it is effective for transferring visual style, this method has limited generalization ability and often overfits to the reference image's specific features, making it difficult to generate entirely new characters that are independent of that reference.

Other multimodal approaches have been developed to guide image generation using a combination of text, image, and style instructions. For example, Zhang et al. [5] introduced a GAN-based model that synthesizes images consistent with multimodal prompts, including stylistic and semantic cues. However, their method struggles to maintain diversity in generated outputs, producing images that are too closely aligned to the conditioning signals. In contrast, our work focuses on testing and understanding the creative capacity of diffusion models when constrained by structured character attributes rather than fixed visual exemplars.

3 Dataset

3.1 Overview

Our dataset consists of 85 unique *Genshin Impact* characters, each represented by an official portrait image that includes the character's name (in both Chinese and English), their vision icon, and a short textual description. In addition, we provide structured character attributes in CSV format containing several categorical attributes for each character: **region**, ***vision***, **weapon type**, **body figure**, and **rarity**, along with three textual attributes, **name**, **constellation** and **affiliation**. The meanings and value ranges of these attributes are explained in the following subsection.

All images were obtained from official artwork released by HoYoverse, resized to 2250×2250 pixels, and the character attribute data were collected from the public dataset on Kaggle.

3.2 Attributes

Table 1: Summary of attributes in character attributes

Attribute	Value Range	Meaning	Type
Name	Any string	English name of the character	Textual
Region	M, L, I, S, F	Character's home region	Categorical
Vision	A, G, D, C, P, H, E	Elemental power the character wields	Categorical
Weapon type	Sw, Bw, Cl, Ca, Po	Primary weapon used in combat	Categorical
Body figure	MM, TM, MF, TF, SF	General body shape of the character	Categorical
Rarity	4, 5	Character rarity (star level)	Categorical
Constellation	Any string	Character's symbolic archetype and fate	Textual
Affiliation	Any string	Group, faction, or organization associated with the character	Textual

Regions — M (Mondstadt), L (Liyue), I (Inazuma), S (Sumeru), F (Fontaine). Visions — A (Anemo), G (Geo), D (Dendro), C (Cryo), P (Pyro), H (Hydro), E (Electro). Weapon types — Sw (Sword), Bw (Bow), Cl (Claymore), Ca (Catalyst), Po (Polearm). Body figures — MM (Medium Male), TM (Tall Male), MF (Medium Female), TF (Tall Female), SF (Short Female).

In the generation process, a new character can be represented by any possible combination of categorical attribute values, whereas the textual attributes, such as name, constellation or affiliation, can take any user-defined string.

3.3 Image–Attributes Relations

All samples follow an anime-inspired art style with similar portrait compositions. While attributes such as *name* and *body figure* can be visually identified in the generated images, other attributes exhibit more implicit correlations between visual and semantic elements. In particular, some attributes influence the character's appearance through subtler cues such as color palette, costume design, or accessory motifs.

For example, a character's *region* strongly affects their clothing style, accessories, and overall aesthetic. Each region in *Genshin Impact* draws inspiration from a real-world culture: Mondstadt from Germany, Liyue from China, Inazuma from Japan, Sumeru from Persia and India, and Fontaine from France. Similarly, the color of the *vision* is closely tied to the dominant palette of the outfit, the color of the character's name text, and the background tone of the portrait. Furthermore, the *constellation* often guides symbolic motifs embedded within the character's costume, appearing as repeated visual elements or thematic icons.

Consequently, a well-trained generative model should be capable of learning and reproducing these cross-modal relationships. Generated characters are therefore expected to exhibit coherent stylistic and semantic correspondences between their visual features and the provided attribute combinations.

3.4 Examples and Observations

Figure 1 shows a subset of characters sharing the same *vision* type (Anemo), illustrating how elemental affinity leads to a consistent green–teal palette. In contrast, Figure 2 presents characters from different regions and vision types, including Sucrose from Mondstadt, Ningguang from Liyue, Raiden Shogun from Inazuma, and Furina from Fontaine, highlighting how both element and region influence each character's color scheme and costume design.

Figure 1: Four characters with Anemo *vision* share a consistent green–teal palette [6].

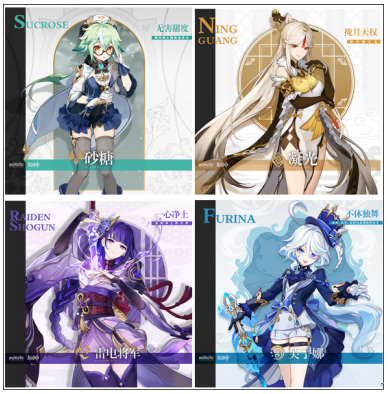

Figure 2: Characters with different *vision* types exhibit distinct color schemes and motifs [6].

4 Model Architecture

4.1 Overview

Our goal is to generate new Genshin–style character images that align with structured character attributes. To achieve this, our architecture combines two complementary components:

1. A **LoRA-refined Stable Diffusion model** that learns the visual *art style* of Genshin characters.

2. A **character attributes Encoder MLP** that learns the semantic *mapping* from character attributes to CLIP's text embedding space.

This design separates the visual and semantic responsibilities: the LoRA module captures the style of Genshin illustrations, while the MLP captures how numeric and textual attributes influence a character's conceptual embedding. By combining both, the system can synthesize visually consistent yet semantically diverse character images.

4.2 LoRA Refinement

The Low-Rank Adaptation (LoRA) module fine-tunes the Stable Diffusion model to reproduce the Genshin visual aesthetic without retraining the entire diffusion network. During LoRA training, we use image–text pairs where all text captions are identical "Genshin-style character". This constant prompt ensures that the LoRA model learns only the shared art style, independent of character-specific semantics.

The refined diffusion model thus focuses solely on the visual rendering characteristics of Genshin art, including line quality, coloring, shading, and composition, while ignoring content details such as region or *vision* type. This LoRA-enhanced model will later serve as the image generator, taking a semantic embedding as conditioning input to produce the final image.

4.3 MLP character attributes Encoder

The MLP serves as a character-attributes-to-embedding mapper, learning how structured character attributes correspond to CLIP's shared semantic space. To build a unified representation of character attributes, we process these inputs as follows:

- **Discrete attributes** are converted into one-hot vectors and concatenated together to form a structured categorical representation.
- **Textual attributes** are encoded using CLIP's frozen text encoder, which transforms text descriptions (e.g., constellation names) into semantically meaningful embeddings.

The one-hot vectors and CLIP text embeddings are then concatenated into a single combined vector, denoted as x_{input}:

$$x_{\text{input}} = [x_{\text{one-hot}}, E_{\text{textattr}}], \tag{1}$$

where $x_{\text{one-hot}}$ represents all concatenated discrete attributes and E_{textattr} represents the CLIP-encoded textual attributes. Before concatenation, all components are normalized to ensure comparable magnitudes.

The combined input vector is passed through the MLP, which predicts an offset vector in CLIP's text embedding space. Each fully connected layer in the MLP is parameterized by a weight matrix W and a bias vector b, and computes

$$h_i = \sigma(W_i h_{i-1} + b_i), \tag{2}$$

where $\sigma(\cdot)$ denotes the activation function (for example, ReLU). Through backpropagation, these weights and biases are optimized so that the sum of the base text embedding and the predicted offset aligns with the CLIP image embedding of the corresponding character.

Let E_{text} denote the CLIP embedding of the prompt "Genshin-style character," and let the MLP output be $E_{\text{offset}} = \text{MLP}(x_{\text{input}})$. The predicted text embedding is therefore

$$E_{\text{pred}} = E_{\text{text}} + E_{\text{offset}}. \tag{3}$$

The training objective minimizes the cosine distance between E_{pred} and the CLIP image embedding E_{img}:

$$L = 1 - \cos(E_{\text{pred}}, E_{\text{img}}), \tag{4}$$

where $\cos(\cdot, \cdot)$ measures angular similarity in CLIP's embedding space. This loss encourages the MLP to learn offsets that adjust the base text embedding toward the semantic direction of the image embedding.

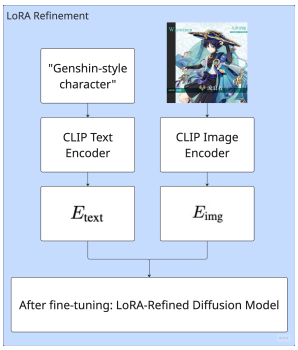

Figure 3: LoRA fine-tuning process. Each image–caption pair trains LoRA to reproduce the shared art style.

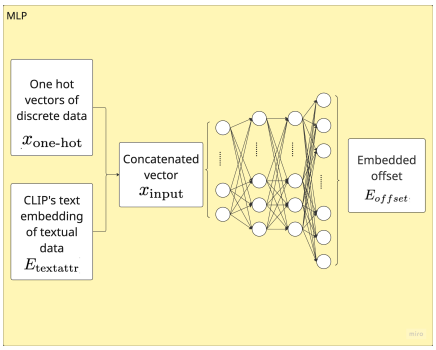

Figure 4: MLP character encoder. Encodes attributes into CLIP's text embedding space.

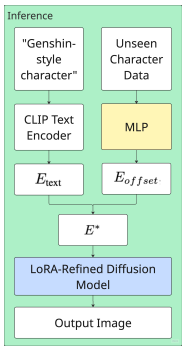

Figure 5: Unified model combining LoRA and MLP encoder to generate Genshin-style characters.

4.4 Inference Stage

During inference, we encode the base text prompt *"Genshin-style character"* using CLIP's text encoder to obtain E_{text}. For a new (unseen) set of character attributes, we use the trained MLP to

compute a semantic offset $E_{offset} = f_\theta(c)$, where c represents the input character attributes. These embeddings are combined as:

$$E^* = E_{\text{text}} + E_{offset}, \tag{5}$$

and E^* is provided to the frozen LoRA-refined diffusion model as the conditioning vector. The diffusion process denoises random noise into a latent representation that, when decoded, becomes a new Genshin-style character image consistent with the provided character attributes.

5 Ethical Considerations

Our project raises a few ethical concerns about using generative models in creative fields, especially when the data come from copyrighted or artist-made materials.

5.1 Copyright and Data Consent

We fine-tuned our model using character images from HoYoverse's *Genshin Impact* without asking for permission from the company or the artists. Even though this work is for research only and not for commercial use, this still raises questions about ownership and consent. Right now, there are no clear rules on whether using copyrighted art for model training counts as fair use. Because of that, it's hard to say where the line is between learning a style and copying it too closely.

5.2 Bias and Representation

If the training dataset of the stable diffusion model we fine-tuned is biased, the model could repeat or even exaggerate those biases in its outputs. For example, if certain races or styles are underrepresented, the model might generate stereotypical results. While we have no control over the composition of the original training dataset, it is important to be aware of these biases and interpret generated results critically, especially in creative or representational contexts.

5.3 Impact on Artists

As AI tools get more advanced, companies might use them to cut costs by generating art instead of hiring artists. While our goal is to study creativity, not replace people, this could still affect how artists are valued or hired in the future.

5.4 Long-Term Effects on Creativity

Recently, work in [7] found that using AI too much can make people's creativity converge, meaning artists might start to create in similar ways if they use AI outputs as references. If this trend continues, human creativity could slowly become more limited instead of more diverse. This is an important question for the long term: how do we make sure AI supports, rather than narrows, human creativity?

6 Work Division

The project is divided into three components: LoRA fine-tuning, MLP stat encoding, and system integration. Each member focuses on one major component, with coordinated milestones to align dependencies. Table 2 summarizes responsibilities and weekly progress.

Mary is responsible for LoRA fine-tuning of the Stable Diffusion model, including minimal dataset preparation, caption pairing, training configuration, and visual quality evaluation.

Lawrence implements and trains the MLP stat encoder, covering data preprocessing, CLIP embedding alignment, network design, and evaluation using cosine similarity loss.

Xiaohan integrates both components into a unified generation pipeline, manages version control, performs inference and visualization, and leads final report preparation.

All members will contribute equally to report writing, focusing on their respective sections.

Table 2: Weekly responsibilities and milestones.

Week	Mary (LoRA)	Lawrence (MLP Encoder)	Xiaohan (Integration)
1	Prepare LoRA captions and environment; begin fine-tuning on Genshin dataset.	Preprocess character stats; design MLP input schema.	Set up repository, project structure, and architecture draft.
2	Continue LoRA training; evaluate visual quality; save checkpoints.	Implement MLP and extract CLIP embeddings for training.	Build integration template; test data flow.
3	Fine-tune LoRA hyperparameters; finalize checkpoint.	Train and tune MLP; analyze embedding alignment and loss.	Integrate LoRA and MLP outputs; run initial inference.
4	Support final inference tests; analyze generated styles.	Summarize MLP performance and assist integration.	Run full pipeline; prepare visualizations and report figures.

Team meetings will be held twice a week, every **Monday and Thursday at 7:00 PM**, to review progress, coordinate dependencies, and resolve issues promptly. Each member is expected to commit approximately 8–10 hours per week, including experimentation, documentation, and discussions.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. Stable Diffusion paper.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

[4] Kai Zhang, Jianming Yang, David Lee, and Yizhou Zhao. Style-aligned image generation via shared attention. *arXiv preprint arXiv:2403.09855*, 2024.

[5] Yue Zhang, Jianyu Wang, Lei Chen, and Zhi Liu. Multimodal-gan: Learning cross-modal relations for text-, image-, and style-guided generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 20(3):1–18, 2024. doi: 10.1145/3757749.3757753.

[6] HoYoverse. Official character artwork from *Genshin Impact*. `https://genshin.hoyoverse.com/en/`, 2020. Accessed November 2025.

[7] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking, 2024. Submitted 24 Sep 2024; revised 15 Feb 2025.