Genshin-Style Character Generation Using Diffusion Model

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

Abstract

We investigate how a generative model handles variation 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.

GitHub Repository

1 Introduction

Audiences remain interested in an IP when new characters or events add variation without breaking the established world. This expectation creates a challenge for artistic producers because they must determine which elements can change and which must remain fixed. As AI tools enter illustration pipelines, the ability of generative models to respect these constraints becomes more important. Studios such as HoYoverse, known for the distinct style of Genshin Impact, would need models that generate content consistent with the IP's visual structure if they adopt AI-assisted workflows. This leads to a central question: to what extent can a model identify both explicit and implicit constraints while introducing novelty in a controlled way?

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 propose 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. To evaluate this, we measure both the model's ability to follow attribute-specific constraints and its consistency with the established Genshin Impact visual style.

*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 72 unique *Genshin Impact* characters, each represented by an official portrait image, and structured character attributes in CSV format. The details of the attributes are explained in the following subsection.

All images were obtained from official artwork released by HoYoverse, resized to 512×512 pixels, and the character attribute data were collected from genshin-impact fandom.

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
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, corresponding to Germany), L (Liyue, China), I (Inazuma, Japan), S (Sumeru, Persia and India), F (Fontaine, France).
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 *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* strodatangly affects their clothing style, accessories, and overall aesthetic. And the color of the *vision* is closely tied to the dominant palette of the outfit, 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].

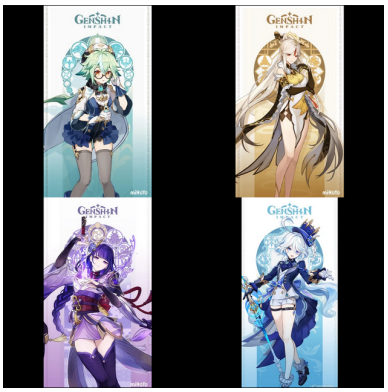

Figure 2: Characters with different *vision* types exhibit distinct color schemes and motifs [6].

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 whose captions describe the key visual elements of each image. The goal is to help the model learn which elements are tied to the prompt and which can vary during generation. 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.

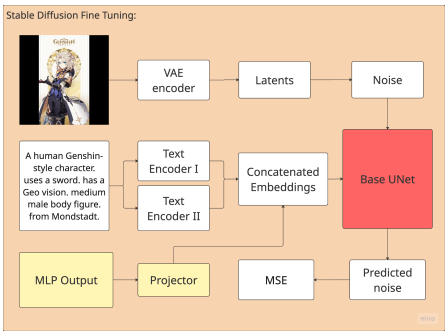

Figure 3: LoRA and projector fine-tuning process. Each image–caption pair trains the stable diffusion model to reproduce the shared art style.

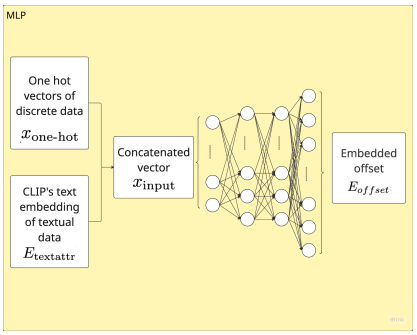

Figure 4: MLP character encoder. Encodes attributes into CLIP's text embedding space.

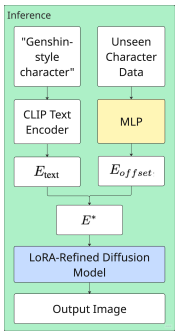

Figure 5: Unified model combining LoRA and MLP encoder to generate Genshin-style characters.

4.4 Projector

The MLP produces an embedding in CLIP space, which has dimensionality $(b, 512)$. However, Stable Diffusion XL does not operate on CLIP embeddings directly. Its text-conditioning is represented by two text encoders whose concatenated hidden states form a $(b, 77, 2048)$ representation. Because the SDXL UNet was trained exclusively on this internal embedding space, a raw CLIP embedding is both shape-incompatible and semantically incompatible with what the diffusion model expects.

To bridge this gap, we adopt an idea inspired by Textual Inversion [7]: instead of learning a new token embedding directly, we learn a projector function that maps the CLIP-space vector into the SDXL embedding space.

The projector is a small MLP that takes a $[1, 512]$ CLIP embedding and outputs a vector in $[1, 2048]$. During inference, this vector is injected by replacing the final token of the concatenated SDXL text embedding. This design makes the projected vector behave like a meaningful "semantic token" that the UNet can use during its cross-attention operations.

By learning the projector end-to-end during fine-tuning, the system discovers how to translate high-level semantic attributes encoded by CLIP into the latent "language" of SDXL. Thus, instead of learning a single static embedding (as in textual inversion), we learn a mapping function that can generate many embeddings conditioned on arbitrary character metadata.

4.5 Inference Stage

During inference, we first encode a detailed descriptive prompt using SDXL's two text encoders to obtain a pair of embeddings that already lie in the Stable Diffusion embedding space. These embeddings are concatenated to form the base conditioning vector.

Separately, the character attribute vector is provided to the MLP, which produces an offset embedding in CLIP space. This offset is then projected into the Stable Diffusion space and used to replace the final column of the concatenated conditioning vector. The resulting conditioning vector is supplied to the LoRA-refined diffusion model, which denoises random latent noise into a new Genshin-style character image consistent with the specified attributes.

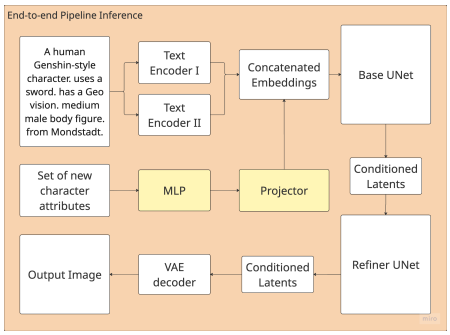

Figure 6: The end to end training process of our pipeline.

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-assisted drawing tools become more capable, studios may integrate them into production pipelines to reduce costs or speed up asset creation. While our work aims to understand controllable generation rather than replace artists, such tools could still change how artistic labor is valued, distributed, or hired in professional settings. In addition, work in [8] 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 Results

Figure 7:
name = "Albeless"
constellation = "Fox"
affiliation = "Knights of Favonius"
region = "Mondstadt"
vision = "Electro"
weapon = "Claymore"
body = "Medium Male"
prompt = "A human Genshin-style character, uses a claymore, has a Electro vision, medium male body figure, from Mondstadt."

Figure 8:
name = "Albeless"
constellation = "Banana"
affiliation = "Arataki Gang"
region = "Inazuma"
vision = "Dendro"
weapon = "Bow"
body = "Short Female"
prompt = "A human Genshin-style character. uses a bow. has a Dendro vision. short female body figure."

Figure 9:
name = "Albeless"
constellation = "Banana"
affiliation = "Arataki Gang"
region = "Inazuma"
vision = "cyro"
weapon = "Bow"
body = "Short Female"
prompt = "A human Genshin-style character. uses a bow. has a cyro vision. medium female body figure, from Inazuma"

These examples show how the model behaves under different attribute configurations.

The model successfully captures several coarse-level attributes specified in the prompt. It produces imagery that is stylistically consistent with Genshin Impact, and in some cases it reliably expresses the vision through color motifs (Figure 7-8), while some fails (Figure 9). Gender and body type are mostly correctly reflected, and the model is able to synthesize a character that appears novel rather than a copy of any training example, while generating from similar prompt and embeddings ends up with similar output(Figure 8-9). In addition, some elements of the game's visual language, such as the Genshin logo and circular background ring, are reproduced consistently for all outputs.

However, the model struggles with finer-grained semantic constraints. Region-specific outfit design, such as Mondstadt's or Inazuma's characteristic tracht-inspired clothing, is not captured. Affiliation (e.g., Knights of Favonius) also fails to appear in any recognizable visual form. Although the generated hairstyle loosely resembles the intended constellation theme for Figure 7, this appears to be incidental, as the model produces substantially different appearances across runs and it failed completely for Figure 8-9. These inconsistencies indicate that the system has difficulty binding highly specific symbolic attributes to stable visual realizations, highlighting a limitation in the projector-based conditioning and the underlying diffusion model's capacity for fine-grained semantic control. That is to say, the model often defaults to generic stylistic patterns instead of encoding the specifically requested semantics.

To complement the qualitative analysis of the generated images, we also evaluate the MLP that predicts the semantic offset, in order to assess how well it learns the attribute–embedding relationship used for conditioning. The MLP achieved a maximum validation accuracy of 0.8492 as measured by cosine similarity. The MLP achieved this accuracy after training for 40 epochs, and further training did not increase the accuracy or decrease the loss value. This could be the result of limited training data, as only 73 image-text pairs were present in the training dataset. Additionally, when provided with text embeddings of characters with particular attributes such as the sentence "A Pyro character

from Mondstadt", the MLP evaluated some characters as being close to the description even though they lacked the required attributes.

7 Discussion

To compare the effects of each component, we present ablations of the diffusion model with and without LoRA and the projector MLP. The baseline SDXL output is included for reference and behaves as expected, showing no Genshin-specific style because it was not trained for this domain.

Among all components, LoRA contributes the majority of the controllable semantics. LoRA directly adapts the diffusion UNet, so it learns to bind visual tokens (e.g., “Electro vision,” “medium male body”) to consistent image-space patterns. This is where most coarse-level attribute fidelity comes from.

By contrast, the projector MLP provides a minor conditioning signal. It operates in a much smaller capacity regime, and it must translate symbolic character attributes into an embedding space that was never trained for such structured semantic composition. As a result, the MLP mainly helps the model stay “on theme,” but it cannot reliably enforce specific outfit motifs, faction details, or region aesthetics. The diffusion model ultimately listens far more to the adapted UNet pathways (LoRA) than to the projector.

Meaningful control in our system mostly comes from modifying the diffusion model itself through LoRA. The small projector MLP doesn’t have enough capacity to reliably attach detailed symbolic attributes to visual features, so it only provides loose thematic guidance. A stronger conditioning path—such as using a larger projector, adding cross-attention modulation, or training with more explicit attribute labels—would likely help the model connect fine-grained attributes to consistent visual outputs.

Figure 10: base diffusion

Figure 11: base+LoRA

Figure 12: base+MLP

Figure 13: base+LoRA+MLP

8 Limitations

The generated images cannot be directly used without refinement. We observe several limitations, along with possible approaches to alleviate them:

8.1 Incorrect face and hand structures

The model often fails to draw faces and hands correctly under our constraint-driven prompts. Increasing the domain strength in the prompt (for example, adding “highly detailed, photorealistic”) helps the model produce anatomically correct human structures. However, this also suppresses how strongly our custom constraints (e.g., region, weapon, vision) influence the generation, creating a trade-off. We suspect this arises because LoRA + projector training modifies only a small subset of the UNet parameters, so the model cannot fully integrate complex structural information from the constraints. A possible extension is to jointly fine-tune deeper UNet blocks or train with stronger reconstruction losses that emphasize hands and facial geometry.

8.2 More noise and artifacts in larger resolutions

Higher-resolution outputs tend to contain noodle-like artifacts. This is likely due to limited exposure to high-resolution data during LoRA training and the projector not being trained across multiple scales

of the latent space. One possible extension is to apply a super-resolution or face-restoration module after sampling, or to train a second-stage LoRA specifically tuned for high-resolution refinement.

8.3 Only partial adherence to semantic constraints

While the model reliably follows coarse rules—such as matching background color to elemental vision—it fails to capture more subtle constraints like nation-specific outfit patterns. We suspect this occurs because the projector injects only a single token's worth of information, and SDXL has learned no direct correlation between those embeddings and the fine-grained clothing styles. One possible extension is to replace the single-token projector with a multi-token projector, or to condition SDXL through cross-attention adapters that provide richer attribute signals.

9 Conclusion

Overall, the system can generate characters that match the intended style and several of the semantic constraints, but it still fails on anatomy and detailed visual rules that trained artists would consider essential. Increasing prompt detail improves structure but suppresses the constraints provided by the MLP–projector, revealing a tension between style accuracy, semantic control, and visual correctness. Our results show that constraint-guided generation is feasible but incomplete, and future work should focus on architectures that integrate structured conditioning more natively rather than treating it as an external signal.

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] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[8] 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.