Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

Interactive Visual Learning for Stable Diffusion

Seongmin Lee[1] , **Benjamin Hoover**[1,2] , **Hendrik Strobelt**[2] , **Zijie J. Wang**[1] ,
ShengYun Peng[1] , **Austin Wright**[1] , **Kevin Li**[1] , **Haekyu Park**[1] ,
Haoyang Yang[1] and **Duen Horng Chau**[1]
[1]Georgia Tech
[2]IBM Research
{seongmin, bhoov}@gatech.edu, hendrik.strobelt@ibm.com,
{jayw,speng65,apwright,kevin.li,haekyu,alexanderyang,polo}@gatech.edu

Abstract

Diffusion-based generative models' impressive ability to create convincing images has garnered global attention. However, their complex internal structures and operations often pose challenges for non-experts to grasp. We introduce Diffusion Explainer, the first interactive visualization tool designed to elucidate how Stable Diffusion transforms text prompts into images. It tightly integrates a visual overview of Stable Diffusion's complex components with detailed explanations of their underlying operations. This integration enables users to fluidly transition between multiple levels of abstraction through animations and interactive elements. Offering real-time hands-on experience, Diffusion Explainer allows users to adjust Stable Diffusion's hyperparameters and prompts without the need for installation or specialized hardware. Accessible via users' web browsers, Diffusion Explainer is making significant strides in democratizing AI education, fostering broader public access. More than 7,200 users spanning 113 countries have used our open-sourced tool at https://poloclub.github.io/diffusion-explainer/. A video demo is available at https://youtu.be/MbkIADZjPnA.

1 Introduction

Diffusion-based generative models [Rombach *et al.*, 2022] like Stable Diffusion [Stability AI, 2022] have captured global attention for their impressive image creation abilities, from AI developers, designers, to policymakers. However, the popularity and progress of generative AI models have sparked ethical [Brusseau, 2022] and social concerns, such as accusations of artistic style theft by AI image generators [Sung, 2022; Choudhary, 2022]. Policymakers are also discussing ways to combat malicious data generation and revise copyright policies [Engler, 2023; Ryan-Mosley, 2023; U.S. Copyright Office, 2023]. There is an urgent need for individuals from many different fields to understand how generative AI models function and communicate effectively with AI researchers and developers [Dixit, 2023; Hendrix, 2023].

Key challenges in designing learning tools for Stable Diffusion. Stable Diffusion iteratively refines *noise* into a high-resolution image's vector representation, guided by a text prompt. Internally, the prompt is tokenized and encoded into vector representations by the *CLIP*'s *Text Encoder* [Radford *et al.*, 2021]. With text representations' guidance, Stable Diffusion improves the image quality and adherence to the prompt by incrementally denoising the image's vector representation using the *UNet* [Ronneberger *et al.*, 2015] and the *Scheduler* algorithm [Nichol and Dhariwal, 2021]. The final image representation is upscaled to a high-resolution image. The crux of learning about Stable Diffusion tems from the complex interplay between the multiple neural network subcomponents, their intricate operations, and the iterative nature of image representation refinements. Such complex interactions are challenging even for experts to comprehend [von Platen, 2022]. While some articles [Alammar, 2022] and video lessons [Howard, 2023] explain Stable Diffusion, they often focus on model training and mathematical details.

Contributions. In this demonstration, we contribute:

- **Diffusion Explainer, the first interactive visualization tool designed for non-experts** to learn how Stable Diffusion transforms a text prompt into a high-resolution image (Fig. 1), overcoming design challenges in developing learning tools for Stable Diffusion. Diffusion Explainer integrates an overview of Stable Diffusion's complex structure with explanations of their underlying operations enabling users to fluidly transition between multiple abstraction levels through animations and interactive elements.
- **Real-time interactive visualization** to discover how Stable Diffusion's hyperparameters and text prompt affect image generation, empowering users to experiment with their settings and gain insight into each hyperparameter's impact without the need for complex mathematical derivations.
- **Open-sourced, web-based implementation** that broadens the public's education access to modern generative AI without requiring any installation, advanced computational resources, or coding skills. Diffusion Explainer is open-sourced[1] and available at https://poloclub.github.io/diffusion-explainer/. A video demo is available at https://youtu.be/MbkIADZjPnA. With over 7,200 users across 113 countries, Diffusion Explainer is making significant strides in democratizing AI education.

[1]https://github.com/poloclub/diffusion-explainer

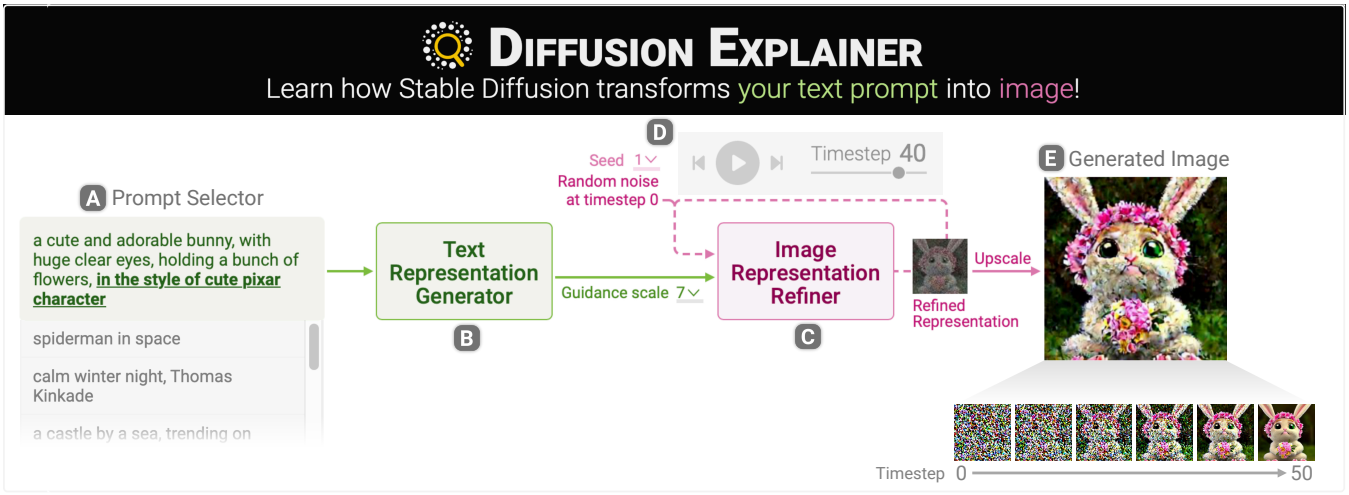

Figure 1: With Diffusion Explainer, users can examine how **(A)** *a text prompt*, e.g., *"a cute and adorable bunny... pixar character"*, is encoded by **(B)** the *Text Representation Generator* into vectors to guide **(C)** the *Image Representation Refiner* to iteratively refine the vector representation of the image being generated. **(D)** The *Timestep Controller* enables users to review the incremental improvements in image quality and adherence to the prompt over timesteps. **(E)** The final image representation is upscaled to a high-resolution image. Diffusion Explainer tightly integrates a visual overview of Stable Diffusion's complex components with detailed explanations, enabling users to fluidly transition between abstraction levels through animations and interactive elements (see Fig. 2 and Fig. 3).

2 System Design and Implementation

Diffusion Explainer is an interactive visualization tool that explains how Stable Diffusion generates a high-resolution image from a text prompt. It incorporates an animation of random noise gradually refined and a *Timestep Controller* (Fig. 1D) that enables users to visit each refinement timestep. From the *Prompt Selector* (Fig. 1A), users select one out of the 13 prompts that follow a template and contain popular keywords identified from literature [Smith, 2022]. Diffusion Explainer provides an overview of Stable Diffusion's architecture, which can be expanded into details via user interactions (Fig. 2, Fig. 3). While users can interactively change Stable Diffusion's two key hyperparameters, guidance scale and random seed, we fix the number of timesteps as 50, a commonly chosen value, and consistently use the Linear Multistep Scheduler [Karras *et al.*, 2022], a fundamental and widely adopted scheduling method. Diffusion Explainer is implemented using a standard web technology stack (HTML, CSS, JavaScript) and the D3.js [Bostock *et al.*, 2011] visualization library.

2.1 Text Representation Generator

The *Text Representation Generator* converts text prompts into vector representations. Clicking on the *Text Representation Generator* expands to the *Text Operation View* (Fig. 2A) that explains how the Tokenizer splits the prompt into tokens and how the Text Encoder encodes the tokens into vector representations. Clicking on the Text Encoder displays the *Text-image Linkage Explanation* (Fig. 2B), which visually explains how Stable Diffusion connects text and image by utilizing the CLIP [Radford *et al.*, 2021] text encoder to generate text representations with image-related information.

2.2 Image Representation Refiner

The *Image Representation Refiner* (Fig. 3) refines random noise into the vector representation of a high-resolution image that adheres to the text prompt. Diffusion Explainer visualizes the image representation of each refinement step in two ways: (1) decoding it as a small image using linear operations [Turner, 2022] and (2) upscaling it to the Stable Diffusion's output resolution (Fig. 1E). Users expands the Image Representation Refiner to access the *Image Operation View* (Fig. 3A), which explains how the UNet neural network [Ronneberger *et al.*, 2015] predicts the noise to be removed from the image representation.

The guidance scale hyperparameter, which controls the image's adherence strength to the text prompt, is described at the bottom, and further explained in the *Interactive Guidance Explanation* (Fig. 3B). Using a slider, users can experiment with different guidance scale values to better understand how higher values lead to stronger adherence of the generated image to the text prompt.

3 Demonstrating Diffusion Explainer

We provide a demonstration scenario both to illustrate how people with limited experience with Stable Diffusion may benefit from using Diffusion Explainer and to describe what we will show the audience.

3.1 Demonstration Scenario

Troy, a government policymaker overseeing AI image creation in the entertainment and media industries, has recently received concerns from artists. They express worry that their artwork has been exploited by AI models to create commercial products without their consent [AMELION, 2023]. Troy is eager to help these artists in getting compensated for their

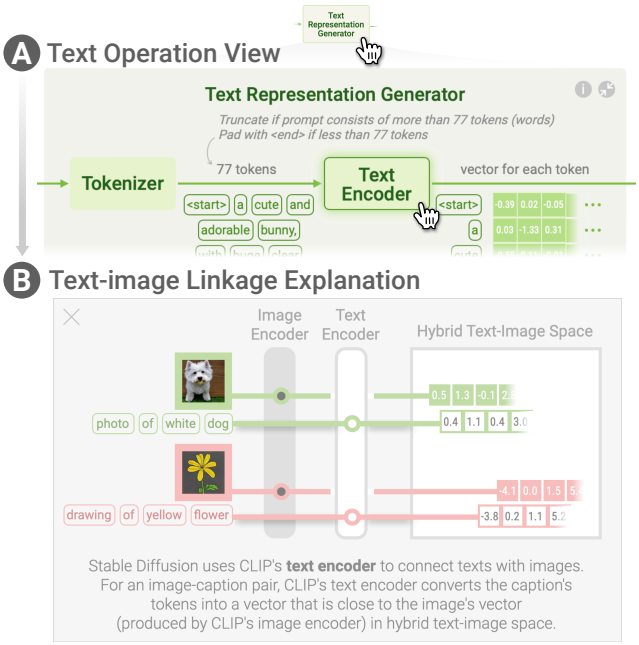

Figure 2: To understand how Stable Diffusion converts a text prompt into vector representations, users click on the *Text Representation Generator*, which smoothly expands to **(A)** the *Text Operation View* that explains how the prompt is split into tokens and encoded into vector representations. **(B)** The *Text-image Linkage Explanation* demonstrates how Stable Diffusion bridges text and image, enabling text representations to guide the image generation process.

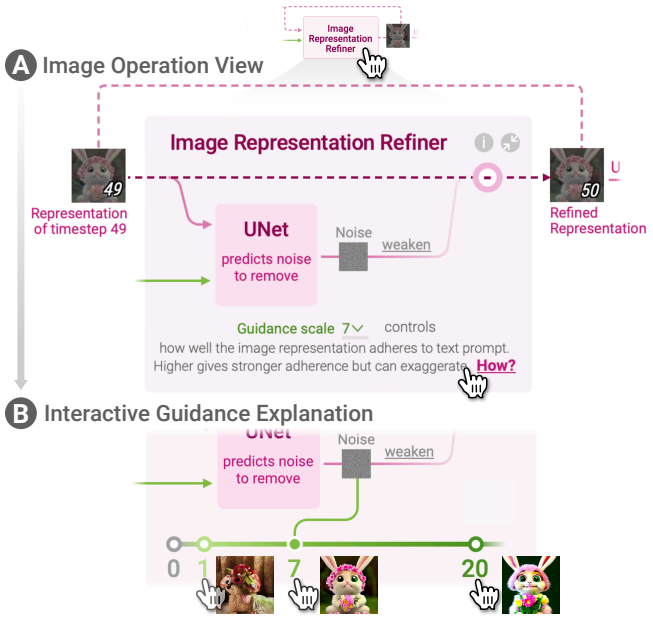

Figure 3: Users learn how Stable Diffusion gradually refines noise into a high-resolution image's vector representation aligned with the text prompt by selecting the *Image Representation Refiner* from the high-level overview. This smoothly expands to **(A)** the *Image Operation View* that demonstrates how the noise is iteratively predicted and removed from the image representation. **(B)** The *Interactive Guidance Explanation* enables users to interactively experiment with various guidance scale values (0, 1, 7, 20) to better understand how higher values lead to stronger adherence.

contributions. He has found a tool that could potentially address their concerns, which would attribute AI-generated images to human artists [Huber and Troynikov, 2023; anton, 2023]. However, before proposing any policies, Troy needs to understand how and if such attribution may work.

Troy launches Diffusion Explainer which illustrates how Stable Diffusion transforms a text prompt into a high-resolution image through an iterative process (Fig. 1). He identifies two controllable hyperparameters: *random seed* and *guidance scale*. Adjusting the random seed from 1 to 2 and 3, Troy observes substantial changes in the generated image. Intrigued by these variations, he examines timestep 1 using the *Timestep Controller* (Fig. 1D) and discovers that different random seeds yield different initial noises, thus generating diverse images. Continuing his exploration, Troy experiments with different guidance scale values. He notes that a guidance scale value of 7 produces a realistic image closely aligned with the text prompt, while values of 1 or 20 result in images that are hard to interpret or exaggerated.

To delve into the details of how the text prompt is processed, Troy clicks on the *Text Representation Generator* to expand it into the *Text Operation View* (Fig. 2A). Here, he discovers that the prompt is tokenized and converted into vector representations. Seeking clarity on how text is connected to the image, he then displays the *Text-image Linkage Explanation* (Fig. 2B) and learns that Stable Diffusion's text representations contain image-related information. Troy proceeds to understand the refinement of image representation by examining the *Image Operation View* (Fig. 3A). He discovers that each refinement step involves UNet's noise prediction and removal, with the guidance scale hyperparameter controlling the adherence of the generated image to the prompt. Intrigued, Troy accesses the *Interactive Guidance Explanation* (Fig. 3B) and learns that the model predicts two types of noise, each of which is generic and prompt-specific. The final noise is a weighted sum of these noises, with the weight being controlled by the guidance scale.

With an improved understanding of the image generation process of Stable Diffusion, Troy recognizes that image analysis alone, without considering text prompts, will not suffice to discern how an artist's creations contributed to AI-generated images. He asserts that further research is imperative to accurately attribute AI-generated images.

4 Conclusion

Diffusion Explainer, the first interactive visualization for non-experts, explains how Stable Diffusion generates high-resolution images from text prompts. It tightly integrates a visual overview of Stable Diffusion's components with detailed explanations of their underlying operations and provides real-time hands-on experience to change Stable Diffusion's hyperparameters and prompts on the browser without any installation or hardware requirements. More than 7,200 users spanning 113 countries have used Diffusion Explainer.

References

[Alammar, 2022] Jay Alammar. The illustrated Stable Diffusion. https://jalammar.github.io/illustrated-stable-diffusion/, 2022. Accessed on: 2023-04-30.

[AMELION, 2023] AMELION. https://twitter.com/amelion_/status/1651193228677218304, 2023. Accessed on: 2023-04-26.

[anton, 2023] anton. Announcing Stable Attribution - A tool which lets anyone find the human creators behind AI generated images. https://twitter.com/atroyn/status/1622355473193381888, 2023. Accessed on: 2023-04-30.

[Bostock *et al.*, 2011] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 Data-driven Documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[Brusseau, 2022] James Brusseau. Acceleration AI Ethics, the Debate between Innovation and Safety, and Stability AI's Diffusion versus OpenAI's Dall-E. *arXiv preprint arXiv:2212.01834*, 2022.

[Choudhary, 2022] Lokesh Choudhary. Stable Diffusion is Now Accused of 'Stealing' Artwork. https://analyticsindiamag.com/stable-diffusion-is-now-accused-of-stealing-artwork/, 2022. Accessed on: 2023-04-30.

[Dixit, 2023] Pranav Dixit. Meet The Three Artists Behind A Landmark Lawsuit Against AI Art Generators. https://www.buzzfeednews.com/article/pranavdixit/ai-art-generators-lawsuit-stable-diffusion-midjourney, 2023. Accessed on: 2023-04-30.

[Engler, 2023] Alex Engler. Early thoughts on regulating generative AI like ChatGPT. *Brookings Institution*, 2023. Accessed on: 2023-04-30.

[Hendrix, 2023] Justin Hendrix. Generative AI, Section 230 and Liability: Assessing the Questions. *Tech Policy Press*, 2023. Accessed on: 2023-04-30.

[Howard, 2023] Jeremy Howard. From Deep Learning Foundations to Stable Diffusion. https://www.fast.ai/posts/part2-2023.html, 2023. Accessed on: 2023-04-30.

[Huber and Troynikov, 2023] Jeff Huber and Anton Troynikov. Stable Attribution. https://www.stableattribution.com, 2023. Accessed on: 2023-04-30.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rombach *et al.*, 2022] 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*, pages 10684–10695, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Ryan-Mosley, 2023] Tate Ryan-Mosley. An early guide to policymaking on generative AI. *MIT Technology Review*, 2023. Accessed on: 2023-04-30.

[Smith, 2022] Ethan Smith. A Traveler's Guide to the Latent Space. https://sweet-hall-e72.notion.site/A-Traveler-s-Guide-to-the-Latent-Space-85efba7e5e6a40e5bd3cae980f30235f#976ba690a0904431aac693d59830a92c, 2022. Accessed on: 2023-04-29.

[Stability AI, 2022] Stability AI. Stable Diffusion Public Release. https://stability.ai/blog/stable-diffusion-public-release, 2022. Accessed on: 2022-08-22.

[Sung, 2022] Morgan Sung. Lensa, the AI portrait app, has soared in popularity. But many artists question the ethics of AI art. *NBC News*, 2022. Accessed on: 2023-04-30.

[Turner, 2022] Kevin Turner. Decoding latents to RGB without upscaling. https://discuss.huggingface.co/t/decoding-latents-to-rgb-without-upscaling/23204/2, 2022. Accessed on: 2023-04-30.

[U.S. Copyright Office, 2023] U.S. Copyright Office. Copyright Office Launches New Artificial Intelligence Initiative. https://www.copyright.gov/newsnet/2023/1004.html, 2023. Accessed on: 2023-04-30.

[von Platen, 2022] Patrick von Platen. Testing Stable Diffusion is hard. https://github.com/huggingface/diffusers/issues/937, 2022. Accessed on: 2023-04-30.