EuroVis Workshop on Visual Analytics (2025)
H.-J. Schulz and A. Vilanova (Editors)

Integrating Layer-Wise Relevance Propagation with Stable Diffusion for Enhanced Interpretability

Christian Auman[1], Deepshikha Bhati[1], Kyle Arquilla[1], Fnu Neha[1], and Angela Guercio[1]

[1]Kent State University, Kent, Ohio, USA

Abstract

Diffusion-based generative models, such as Stable Diffusion and DALL-E, have revolutionized artificial intelligence by enabling high-quality image generation from textual descriptions. Despite their success, these models raise ethical concerns, such as style appropriation and misuse, closely tied to the interpretability and transparency of the underlying mechanisms. This paper introduces a framework integrating Layer-wise Relevance Propagation (LRP) into the Stable Diffusion model to enhance interpretability. LRP assigns relevance scores to specific elements of textual prompts, allowing users to understand and visualize how input text influences image generation. We also present an interactive web-based visualization tool that supports intuitive exploration of diffusion processes. By improving interpretability, this approach fosters responsible use of generative AI technologies. A user study involving 35 participants demonstrates the tool's accessibility and effectiveness.

CCS Concepts

• ***Computing methodologies*** → *Generative AI; Diffusion Models; Stable Diffusion; Layer-wise Relevance Propagation; AI Transparency;* • ***Human-centered computing*** → *Visual analytics;*

1 Introduction

Diffusion-based generative models like Stable Diffusion and DALL-E have become powerful tools for artificial intelligence (AI)-driven image synthesis, creating high-quality visuals from text descriptions. They are widely used in entertainment, creative design, and AI research, supporting applications in concept art, marketing, and automated content creation [GWT22, LHS*24, RDN*22].

However, despite their success, diffusion models introduce challenges related to ethical AI use, bias, and interoperability. These ethical concerns are closely tied to the underlying mechanisms of diffusion models, as their complex architectures influence both their capabilities and limitations. Concerns over artistic style appropriation, misinformation, and dataset biases have sparked discussions on regulation and transparency [Ame23, RLRM24, LHS*24]. Legal disputes highlight the need for attribution mechanisms, as artists claim their styles are replicated without consent [And23].

Models like Stable Diffusion reflect these concerns, as their ability to generate structured images from noise is often impaired by issues of interpret-ability and ambiguity. Stable Diffusion refines random noise into structured images using a CLIP-based text encoder and a U-Net denoising network [Ope22, RBL*22]. However, due to its complex architecture, the model struggles with ambiguous prompts and can introduce unintended artifacts [CER22, Jo23, Mos22]. As the adoption of these models grows, so does the demand for interpretability tools that explain how text prompts influence generated images [Bru22, Mag23, Hen23, LHS*24].

This paper presents an interpretability framework leveraging Layer-wise Relevance Propagation (LRP) [MBL*19] to enhance transparency in Stable Diffusion. By assigning relevance scores to text components, LRP enables tracing the influence of specific prompt elements on generated outputs [CLL23, MBL*19].

Our key contributions include:

- An interactive visualization tool that applies LRP to explain Stable Diffusion's text-to-image transformation, accessible via a web-based implementation (`https://sd-lrp-research.vercel.app/`).
- A systematic analysis of prompt structures and their impact on image generation, supported by LRP-based attribution methods.
- Experimental validation with a human study of 35 participants, assessing interpretability visualization effectiveness.
- Provides browser-based access with an intuitive UI, reducing the technical barrier for exploring generative model behavior.

This study enhances interpretability in generative AI by integrating LRP with diffusion-based image generation, fostering transparency and responsible diffusion-based image synthesis.

2 Related Works

Generative AI, particularly diffusion models like Stable Diffusion, has advanced image synthesis, but interpretability remains a challenge. Tools like *Diffusion Explainer* offer high-level insights but lack fine-grained attribution [LHS*24]. Some studies explore interpretability via saliency maps and gradient-based methods [PJL24], yet they remain coarse.

DOI: 10.2312/eurova.20251102

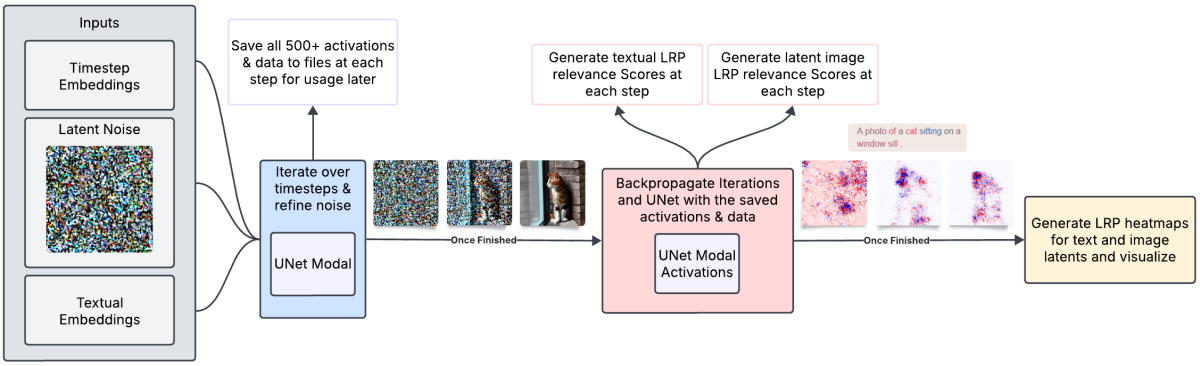

Figure 1: *Overview of the diffusion model interpretability process using LRP. The figure illustrates how Stable Diffusion iteratively refines latent noise using a UNet model while storing activations. Backpropagation is later applied to compute relevance scores, generating LRP heatmaps that highlight influential text and image latent features.*

Layer-wise Relevance Propagation (LRP) provides deeper insight into model behavior [MBL*19, OMS17, SC*20, AGD*24], alongside other attribution techniques like Integrated Gradients and Shapley values [STY17, LL17]. However, their application to diffusion models is still evolving.

While platforms like CNN Explainer [WTS*20] and GAN Lab [KTC*18] aim to simplify complex generative models, they do not provide detailed attribution for diffusion-based architectures, making it difficult to assess how individual inputs influence generated outputs [SK23, HJA20, KP23]. This limitation is particularly relevant for Stable Diffusion, which combines a UNet backbone [RFB15] with CLIP-based encoders [RKH*21] to refine images from noise. The interplay between these components complicates interpretability, especially when dealing with ambiguous prompts or biases present in training data [BP21, YWZ*23, XW23]. To address this, our work integrates LRP into Stable Diffusion, providing fine-grained attribution that reveals how different input components shape the generated output. By mapping relevance scores to specific features, we enhance transparency in generative processes, offering a more interpretable framework for diffusion-based AI models.

3 Design Goals for Explainable AI

Integrating Layer-wise Relevance Propagation (LRP) into Stable Diffusion enhances explainability and user-centered design in generative AI. The following design goals ensure that users can effectively interpret and interact with the model:

1. **Enhanced Transparency:** LRP heatmaps visually convey how individual words in the prompt influence generated images.
2. **Step-by-Step Breakdown:** Users can view changes across time steps, observing how latent representations evolve through the diffusion process.
3. **Interactive Exploration:** The system offers a low-barrier, browser-based interface, minimizing technical entry requirements while acknowledging that interpretability of deep models may still require basic conceptual understanding.
4. **Broad Accessibility:** The interface is intuitive and requires no specialized hardware or installation, ensuring accessibility for users with minimal technical expertise.

By prioritizing interpretability, transparency, and engagement, these design principles advance explainable AI in generative models and empower users to explore and trust AI-generated content more effectively.

4 Interactive Visualization for Stable Diffusion Interpretability

This section outlines the design and implementation of the interactive visualization tool for Stable Diffusion, which is designed to provide an intuitive user experience. The tool allows users to explore the image generation process through key components like the UNet denoising process, text representation, and intermediate activations. By incorporating Layer-wise Relevance Propagation (LRP), the tool enhances interpretability, revealing how textual prompts influence image synthesis and offering deeper insights into the model's decision-making process.

Figure 1 outlines our interpretability framework. The process starts with textual embeddings, timestep embeddings, and latent noise, which are iteratively refined by the UNet denoiser. Over 500 activation states are stored and analyzed, enabling researchers to trace textual influences and identify key activation patterns [SK23, LHS*24]. After image generation, backpropagation-based LRP computes relevance scores, which are then visualized as heatmaps to highlight influential text and latent features. These insights contribute to model fine-tuning, bias mitigation, and improved training strategies in generative AI [GJS23, LHS*24].

Our approach consists of three key components: (1) *Diffusion-Based Image Generation* - Utilizing a pre-trained diffusion model to generate images conditioned on text embeddings, (2) *LRP for Interoperability* - Applying LRP to trace relevance through the model's layers and understand how different regions contribute to image generation, and (3)*Visualization of Relevance Maps* - render-

ing relevance maps to highlight key areas in generated images that influenced the model's output.

The following subsections provide a detailed discussion of each component.

4.1 Diffusion-Based Image Generation

We employ a pre-trained diffusion model to generate images based on text embeddings. The model follows an iterative denoising process, where random latents are progressively refined into coherent images. The process consists of a "forward diffusion" that introduces noise to the latent representation, while a U-net-based network performs the "reverse process" that removes noise step by step reconstructing a meaningful image. The full image generation process is formally described in Algorithm 1.

Algorithm 1 RUN_DIFFUSION_PROCESS

function RUN_DIFFUSION_PROCESS(text_embeddings, unet, timesteps, batch_size, height, width, guidance_scale, scheduler)
    Initialize random latents
    Scale latents using scheduler's noise sigma
    Register hooks in the UNet layers to collect activations
    for each timestep in scheduler.timesteps **do**
        noise_pred, down_block_samples, time_embeddings, weights ← FORWARD_UNET(unet, input_latents, timestep, text_embeddings)
        noise_pred ← APPLY_GUIDANCE(noise_pred, guidance_scale)
        new_latents ← REMOVE_NOISE(input_latents, timestep, noise_pred)
        Save to files: activations, weights, noise_pred, down_block_samples, time_embeddings, and latents for LRP
    end for
end function

4.2 Layer-wise Relevance Propagation (LRP) for Interpretability

To understand how different regions of the generated images contribute to the model's decision-making, we apply Layer-wise Relevance Propagation (LRP). LRP backtraces relevance through the UNet layers by leveraging stored activations and weights from the diffusion process. This technique helps in understanding the model's internal reasoning by attributing relevance scores to specific parts of the generated image.

The LRP process is detailed in Algorithm 2.

4.3 Visualization of Relevance Maps

The final step in our approach is the visualization of computed relevance maps, which enhance interpretability by highlighting key areas of the generated images that significantly influence the model's output. To achieve this, we use Variational Autoencoders (VAEs) to decode the relevance maps into meaningful image representations.

The visualization process is summarized in Algorithm 3.

This structured approach ensures a clear pipeline from diffusion-based image generation to explainability through LRP and visualization.

4.4 Core Features

By integrating the technological components of our interactive visualization tool with real-time controls and a robust backend, the system provides an intuitive and immersive exploration of the image generation process. The system includes the following key components:

Interactive Timestep Controller: This feature lets users navigate through different stages of the diffusion process, visualizing how noise is refined into a clear image at each timestep.

Technology Stack: The frontend is built with React.js and Next.js for a dynamic and interactive user interface, with React Flow used for architectural visualization. The backend is developed using Python and PyTorch to handle deep learning tasks. The model integration utilizes the Hugging Face diffusers library to run the Stable Diffusion model, incorporating LRP to enhance interpretability by highlighting key text and image features.

Algorithm 2 RUN_LRP_PROCESS

function RUN_LRP_PROCESS(num_steps, scheduler, unet, vae, stored_activations, stored_weights, text_embeddings)
    Initialize relevance map *R*, *queries*, *keys*, *values*
    Initialize prev_scores to None
    for each timestep in reverse order **do**
        Load activations, latents, weights, time_embeddings, and donwn_block_samples for current timestep
        if prev_scores is None **then**
            Set prev_scores to last activation
        end if
        Set last activation to prev_scores
        latent_rel_scores, query_scores, key_scores, value_scores ← PROPAGATE_RELEVANCE(unet, vae, layers, activations, down_block_samples, time_embeddings, text_embeddings, latents, weights)
        Normalize latent relevance scores
        Append computed relevance maps for latents, queries, keys, and values
        Set prev_scores to latent_rel_scores
    end for
end function

Algorithm 3 GENERATE_LRP_VISUALIZATIONS

function GENERATE_LRP_VISUALIZATIONS(R, vae, tokenizer, text_input)
    Decode relevance into images using VAE
    for each stored relevance map **do**
        Convert relevance values into visual heatmaps
        Save positive, negative, and combined heatmaps
        Generate text relevance visualizations
    end for
    return generated heatmaps and images
end function

5 Usage Scenario and Visual Walkthrough

This section discusses the experimental results obtained using the proposed framework.

5.1 Interactive Explanations with Time Controller and LRP-Based Interpretability

Consider a user, Alice, who inputs the prompt: *"A photo of a cat sitting on a window sill."* The interactive interface (Figure 2) allows

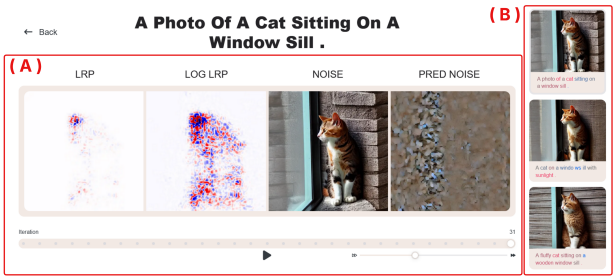

Figure 2: ***Interactive Explanations of Diffusion-Based Image Generation.*** *The time controller visualizes the progressive refinement of an image from Gaussian noise to a structured output, while the gallery view enables a comparative analysis of images generated from prompt variations. LRP-based relevance maps highlight key regions that contribute most to the model's decision-making.*

her to explore the denoising process of a diffusion model using a time controller. The pre-trained UNet model refines random noise into structured images, progressively revealing key attributes like the cat's posture, fur texture, and window sill arrangement.

The left panel visualizes noise evolution at each timestep. The **Noise** view shows raw Gaussian noise, revealing how structured patterns emerge as diffusion progresses. The **Pred Noise** view displays the predicted noise to be removed, helping Alice assess the network's denoising effectiveness. By adjusting the time slider, she can inspect intermediate stages and evaluate the signal preservation.

The left panel (A) includes **LRP** and **Log LRP** visualizations. LRP assigns relevance scores to pixels, highlighting key regions that influence the model's decision. Log LRP emphasizes subtle activations, revealing hidden important features. These heatmaps help Alice understand how the model attends to different image regions.

The right panel (B) also presents a gallery of variations, such as *"A cat on a windowsill with sunlight"* and *"A fluffy cat sitting on a wooden window sill."* Alice can explore how small textual changes affect the generated images, revealing biases in semantic encoding.

By combining Noise analysis, Predicted Noise comparison, LRP-based relevance maps, and interactive gallery exploration, our system improves transparency in diffusion models, providing insights for applications in medical imaging, forensics, and creative design.

5.2 Prompt Influence Analysis Across Diffusion Iterations

A team of scientists explored how a generative model associates textual prompts with visual elements. Using the prompt *"A photo of a cat sitting on a window sill,"* they observed the model's iterative learning process. This progression is illustrated in Figure 3, which corresponds directly to the observations described across Stages A–D.

At **Iteration 2 (Stage A)**, the output was mostly random noise. Heatmaps showed scattered activation patterns, and bar charts revealed very low contributions from all prompt words. Words like *photo* and *window* exhibited slight negative relevance, while *cat* and *sitting* had negligible influence.

Table 1: *User Study Results*

Question	Mean	SD
How easy was it to navigate the site?	4.38	0.98
How easy was it to find the information you wanted?	4.33	0.96
Were the buttons and links clear and understandable?	4.37	1.03
How well did the menu and site layout help you to find information efficiently?	4.37	0.81
How clear were the concepts and terminology?	4.20	0.96
How was your experience with the time-step controller?	4.27	0.98
How well did you understand the purpose of the website within the first few minutes?	4.37	0.81
How clear was the readability of the text and layout?	4.50	0.78
How appealing are the colors on the site?	4.30	0.99
How appealing are the fonts and layout on the site?	4.17	1.05
Overall how visually appealing is the website?	4.10	0.99
How easy was it to understand the plots and graphs?	3.77	1.17
How appropriate were the color representations of the LRP Heatmaps and graphs?	4.30	0.95
How well did the text explanations complement the visuals?	4.37	1.03
Before using this website, how well did you understand Stable Diffusion?	1.97	1.19
After using the website, how well did you understand how Stable Diffusion generates images?	3.83	1.02

By **Iteration 13 (Stage B)**, rough outlines of a cat and a window emerged. The heatmaps showed stronger attention to key regions, and bar charts indicated rising relevance scores for *cat* and *sitting*, though *photo* and *window* remained less influential.

At **Iteration 20 (Stage C)**, the cat's features became more distinct. Positive relevance values for *cat* and *sitting* increased significantly, while *photo* and *window* showed diminished impact, as seen in both the heatmaps and bar charts.

Finally, at **Iteration 30 (Stage D)**, the AI produced a near-complete depiction of the prompt. Heatmaps indicated confident focus, and bar charts showed that *cat* and *sitting* dominated the model's attention, with other terms contributing minimally.

This progression highlights how feedback gradually refines the model's understanding, with heatmaps and bar charts tracking the growing semantic alignment between prompt and output.

6 User Evaluation

We conducted a user study to assess the usability and effectiveness of our proposed interactive tool, targeting non-experts in generative AI. The study included 35 participants (30 students, 5 researchers) from various disciplines with minimal prior experience in generative AI. Participants reported significant improvement in understanding, from a mean of 1.97 before usage to 3.83 after. They highlighted the effectiveness of real-time visual feedback and found the LRP heatmaps especially helpful for associating text and image regions. The goal was to evaluate how well the tool helped users understand the complex concepts behind Stable Diffusion. Participants completed specific tasks and provided feedback on ease of use, design preferences, background understanding, and overall ex-

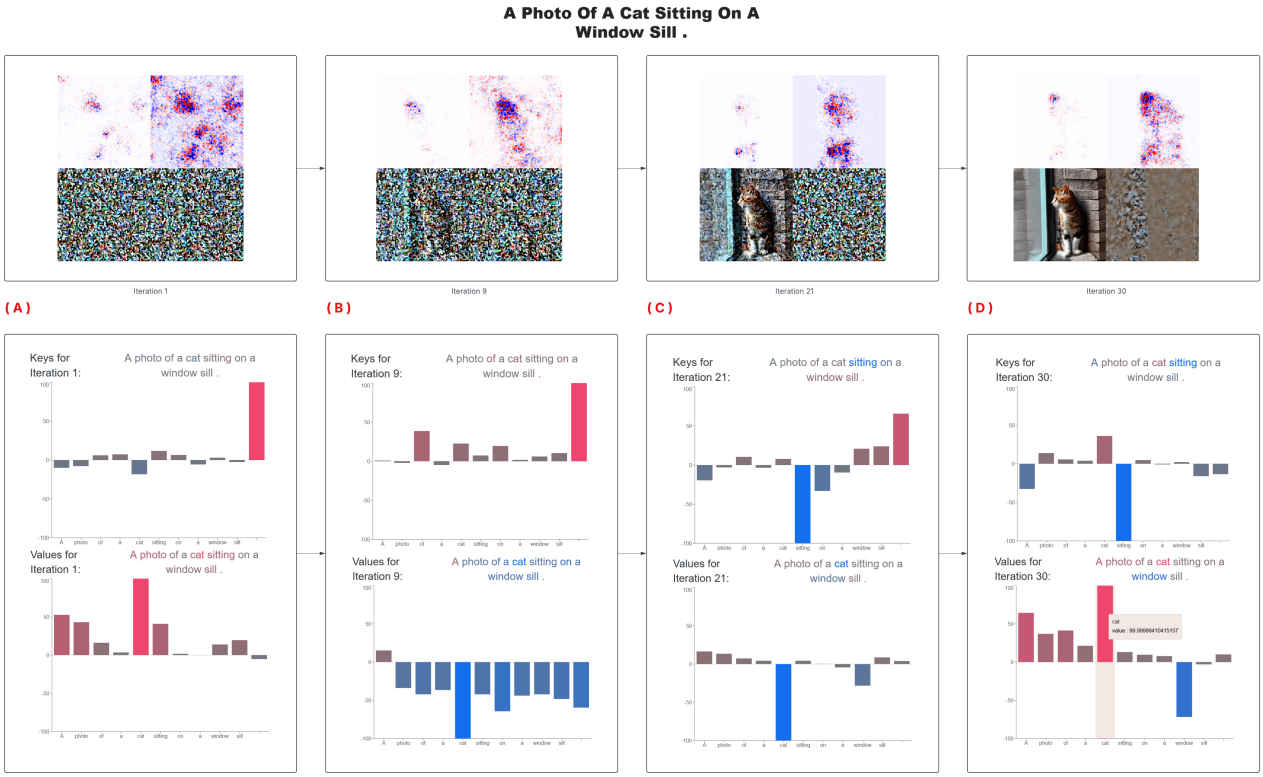

Figure 3: ***The Progressive Evolution of AI Image Generation****. This figure illustrates the step-by-step transformation of an AI-generated image based on the prompt "A photo of a cat sitting on a window sill." (A) In the early iterations, the model struggles to recognize any structure, producing mostly noise. (B) By iteration 13, rough outlines of the cat and window emerge. (C) At iteration 20, the AI refines its attention to key features, enhancing visual clarity. (D) Finally, by iteration 30, the AI successfully generates a clear and structured image that closely matches the prompt. Heatmaps and bar charts indicate how the model progressively improves its focus and word association over time.*

perience. All numbered scores collected during the evaluation were based on a scale of 1-5, with 5 representing the highest rating.

6.1 User Study Results and Insights

The results show in Table 1 that the tool was effective in making complex AI concepts accessible to non-experts. Usability ratings were high, particularly for the intuitive interface (4.37) and ease of navigation (4.38). Interactive features such as real-time visualizations and step-by-step explanations played a key role in helping participants understand Stable Diffusion, leading to a significant improvement in comprehension (mean score of 3.83, up from an initial 1.97). Color representations of the LRP Heatmaps and graphs received positive feedback, with a rating of 4.30.

Participants valued the hands-on learning experience, especially the ability to experiment with different settings and receive immediate feedback. However, some areas for improvement were identified, including simplifying complex plots and enhancing interactivity (e.g., smoother scrolling, dark mode). These insights suggest that further refinements could enhance the tool's accessibility and effectiveness for non-expert users.

7 Conclusion and Future Extensions

In this paper, we presented the development of a web-based interactive visualization system to improve the interpretability of diffusion-based image generation. The system allows users to explore the temporal evolution of images and the semantic relevance of different regions using Layer-wise Relevance Propagation (LRP). With a time controller and gallery view, users can examine how textual prompts influence image generation at different stages.

This tool provides an accessible platform for visualizing the behavior of Stable Diffusion's text-to-image transformation, enabling users to analyze prompt structures and evaluate effectiveness without requiring advanced hardware or technical expertise. Future improvements will include multimodal input support, custom model uploads, educational modules, performance metrics, and expanded LRP visualizations, aiming to enhance the tool's utility for researchers, educators, and students in generative AI.

References

[AGD*24] ARQUILLA K., GAJERA I. D., DARLING M., BHATI D., SINGH A., GUERCIO A.: Exploring fine-grained feature analysis for bird species classification using layer-wise relevance propagation. In *2024 IEEE World AI IoT Congress (AIIoT)* (2024), IEEE, pp. 625–631. 2

[Ame23] AMER S. K.: Ai imagery and the overton window. *arXiv preprint arXiv:2306.00080* (2023). 1

[And23] ANDERSEN K.: Copyright challenges in ai-generated art: Legal and ethical considerations. *AI & Society 38*, 2 (2023), 455–470. 1

[BP21] BIRHANE A., PRABHU V. U.: Multimodal datasets: Misogyny, racism, and harmful stereotypes in vision and language models. *arXiv preprint arXiv:2104.08758* (2021). 2

[Bru22] BRUSSEAU J.: 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). 1

[CER22] CROWSON K., ESSER P., ROMBACH R.: Vqgan-clip: Open-domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583* (2022). 1

[CLL23] CORTIÑAS-LORENZO K., LACEY G.: Toward explainable affective computing: A review. *IEEE Transactions on Neural Networks and Learning Systems* (2023). 1

[GJS23] GOKHALE P., JAIN P., SHRIVASTAVA M.: Explainable ai for generative models: A study on text-to-image synthesis. *Neural Networks 165* (2023), 89–105. 2

[GWT22] GAFNI O., WOLF L., TAIGMAN Y.: Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131* (2022). 1

[Hen23] HENDRIX J.: Generative ai, section 230 and liability: Assessing the questions, 2023. 1

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* (2020). 2

[Jo23] JO T.: *Deep learning foundations*. Springer, 2023. 1

[KP23] KIM D., PARK S.: Latent space interpretability in diffusion-based generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). 2

[KTC*18] KAHNG M., THORAT N., CHAU D. H., VIÉGAS F. B., WATTENBERG M.: Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 310–320. 2

[LHS*24] LEE S., HOOVER B., STROBELT H., WANG Z. J., PENG S., WRIGHT A., LI K., PARK H., YANG H., CHAU D. H. P.: Diffusion explainer: Visual explanation for text-to-image stable diffusion. In *2024 IEEE Visualization and Visual Analytics (VIS)* (2024), IEEE, pp. 96–100. 1, 2

[LL17] LUNDBERG S. M., LEE S.-I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems 30* (2017). 2

[Mag23] MAGAZINE A. I.: Stable diffusion is now accused of stealing artwork, 2023. URL: https://analyticsindiamag.com/ai-news-updates/stable-diffusion-is-now-accused-of-stealing-artwork/. 1

[MBL*19] MONTAVON G., BINDER A., LAPUSCHKIN S., SAMEK W., MÜLLER K.-R.: Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), 193–209. 1, 2

[Mos22] MOSTAQUE E.: Stable diffusion public release. *Stability AI* (2022). 1

[OMS17] OLAH C., MORDVINTSEV A., SCHUBERT L.: Feature visualization. *Distill* (2017). 2

[Ope22] OPENAI: Dall-e 2, 2022. URL: https://openai.com/product/dall-e-2. 1

[PJL24] PARK J.-H., JU Y.-J., LEE S.-W.: Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications 248* (2024), 123231. 1

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695. 1

[RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022). 1

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: 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* (2015), Springer, pp. 234–241. 2

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763. 2

[RLRM24] RAJCIC N., LLANO RODRIGUEZ M. T., MCCORMACK J.: Towards a diffractive analysis of prompt-based generative ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–15. 1

[SC*20] SELVARAJU R. R., COGSWELL M., ET AL.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of IEEE CVPR* (2020). 2

[SK23] SCHRAMOWSKI P., KERSTING K.: Interactive explanations for text-to-image models: Understanding latent space interactions. *arXiv preprint arXiv:2302.04891* (2023). 2

[STY17] SUNDARARAJAN M., TALY A., YAN Q.: Axiomatic attribution for deep networks. In *International conference on machine learning* (2017), PMLR, pp. 3319–3328. 2

[WTS*20] WANG Z. J., TURKO R., SHAIKH O., PARK H., DAS N., HOHMAN F., KAHNG M., CHAU D. H. P.: Cnn explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2020), 1396–1406. 2

[XW23] XU K., WANG L.: Mitigating bias in text-to-image generative models: Techniques and challenges. *IEEE Transactions on Artificial Intelligence* (2023). 2

[YWZ*23] YANG Z., WEI B., ZHANG Y., ET AL.: Evaluating bias and fairness in text-to-image generative models. *arXiv preprint arXiv:2305.07812* (2023). 2