Dynamic Light Video Generation: Extending Spinelli's Interactive Art with Sound Synchronization

Yiyu Chen
University of California, Berkeley
yiyu.chen@berkeley.edu

Yuqin Jiao
University of California, Berkeley
yuqin62@berkeley.edu

Abstract

This project innovatively extends Lucea Spinelli's light-based art by synchronizing vivid, pulsating light trails with musical sound waves, creating an immersive audiovisual experience. Utilizing a two-stage image-to-image custom pipeline, our system dynamically generates light effects that respond to audio inputs, maintaining them within specific regions while preserving the integrity of the background imagery. The process involves extracting musical features to influence initial light trail visualizations, followed by enhancement using the DreamBooth model, which was fine-tuned on artist-inspired images as part of our tailored image-to-image transformation pipeline. The resulting system not only enhances the aesthetic and interactive dimensions of digital media but also introduces new possibilities for interactive digital art installations, emphasizing a novel method to interact with music visually.

1. Introduction

Dynamic visualizations synchronized with audio inputs have profound implications in areas ranging from entertainment to interactive art installations. While traditional approaches have focused on direct visual representations of audio features, the integration of advanced image processing techniques, such as those explored in denoising diffusion probabilistic models [2], significantly enhances interactivity and aesthetic appeal. Inspired by Lucea Spinelli's light-based artworks Fig. 1, our project seeks to transform the way light video generation interact with soundtrack by introducing a novel audiovisual experience.

Utilizing the audio feature extraction approaches, the system first abstracts audio features such as beats and volume to create initial visualizations of light trails. Then the core of our system is an innovative custom two-stage image-to-image pipeline, exemplified by the capabilities of latent diffusion models [5]. This pipeline generates images with background and pulsating light trails responding to the fluctuations in musical sound waves. These trails are then refined to maintain their vibrancy within designated areas of the video frame, ensuring that the background imagery undergoes minimal geometric changes.

These light trails are then meticulously refined to maintain their vibrancy within designated areas of the video frame, ensuring minimal geometric changes to the background imagery. This leads to the use of a sophisticated light effect rendering algorithm, designed to keep the background geometry consistent while adjusting for reasonable light effects and brightness, as detailed in Section 3.5.

The remainder of this paper is organized as follows. Section 2 discusses related work in the fields of synthetic object rendering, diffusion models, and foundational rendering algorithm. Section 3 details our methodology, including the design of a custom two-stage image-to-image pipeline and the specifics of the DreamBooth fine-tuning process. Section 4 concludes the paper and outlines potential future directions for this research. Appendix presents the results of our experiments, showcasing the system's ability to generate synchronized light trails.

(a) time lapse inside a train (b) time lapse on a bench

Figure 1. Lucea Spinelli's time lapse photography artwork

2. Related Work

In computer vision field, significant strides have been made in the areas of inpainting and image-to-image translation, each addressing unique challenges associated with image manipulation and enhancement. Our work builds upon and diverges from these foundational technologies to

specifically address the needs for dynamic light effect generation within videos while preserving the integrity of the background geometry.

2.1. Image-to-Image Translation

Image-to-image translation models transform an input image into a new image, aiming to change the image according to certain attributes while retaining the original context. This field has evolved from simple texture transfers to complex attribute adjustments in images, as demonstrated by the latent diffusion models used in our system [5]. However, traditional frameworks, such as those based on Generative Adversarial Networks (GANs), typically do not manage background elements effectively, nor do they support inpainting to alter objects within a mask and adjust the surrounding background. Such models often lack the capability to selectively enhance specific features like brightness or contrast without affecting the overall geometry of the scene.

2.2. Inpainting Techniques

Inpainting, traditionally used for repairing holes or removing unwanted objects from images, relies on surrounding pixel information to fill gaps in a visually plausible manner. Techniques such as those developed in RePaint [3], which utilize denoising diffusion probabilistic models for inpainting, have pushed the boundaries by generating coherent image completions in masked regions without altering the unmasked background. While these models excel at maintaining the static features of the background, they do not address changes in lighting or color dynamics of the background, which are central to our project.

2.3. Rendering and Realism

Our approach also considers the challenge of rendering synthetic objects into real scenes, which involves bridging traditional and image-based graphics to achieve realistic integration. The foundational work by Debevec et al. [1] has been instrumental in developing techniques that combine global illumination with high dynamic range photography on objects, ensuring that synthetic additions to real environments maintain a natural appearance.

3. Method

The method section describes the innovative approaches to address the challenges of synthesizing and refining dynamic light effects in images based on audio input. The process is divided into three key stages: Preprocessing, the Two-Stage Diffusion Model, and light effect rendering. The complete sequence and interactions between these stages, essential for the controlled manipulation and enhancement of the visual output, are shown in Fig. 2. This diagram illustrates how we strategically apply diffusion techniques and targeted enhancements to achieve precise synchronization and high-quality visual effects.

3.1. Preprocessing

The preprocessing stage of our methodology is designed to synchronize audio with visual effects, and it consists of two main components: audio feature extraction and visual mapping of these features onto images.

3.1.1 Audio Feature Extraction

This component involves the computational extraction of audio signals to derive features that control visual parameters in the generated images. RMS energy, an indicator of the audio signal's power, is crucial for dynamically adjusting the visual intensity based on the audio. As volume is one of the most critical features that listeners perceive when they experience audio, especially in music. Concurrently, rhythmic elements are identified through beat detection, which isolates significant beats within the audio track and generates a corresponding beat signal. This signal is then smoothed using a Gaussian filter to ensure transitions in visual effects are fluid.

The combination of RMS and the beat signal is used to create a composite 'brightness' feature, which dictates the intensity and dynamics of the light effects:

$$Brightness = w_{volume} \times RMS + w_{beat} \times Beat$$

3.1.2 Visual Mapping

The visual mapping process in our project involves two distinct methods for adding light effects onto background images, each influenced by the audio-derived *brightness* feature calculated in the preprocessing stage.

Waveform-Based Light Trails: The first approach, Waveform-Based Light Trails, dynamically generates light trails that directly reflect the audio signal's characteristics over the corresponding frame's duration. By smoothing the audio waveform over the frame's period, we establish a visual representation that mirrors the audio as well as model the light trail. To this foundation, random noise is introduced, varying with the time and intensity of the audio to simulate the natural dynamics of light. The resultant visual effect is a set of light trails whose number is modulated by the calculated 'brightness' feature. This modulation means that more intense audio segments—indicated by higher brightness—result in a denser and more vivid display of light trails, thereby visually amplifying the audio's impact. The Waveform-Based Light Trails mapping process is shown in Fig. 4.

Structured Motion Line Visualization: The second method, Structured Motion Line Visualization, employs a

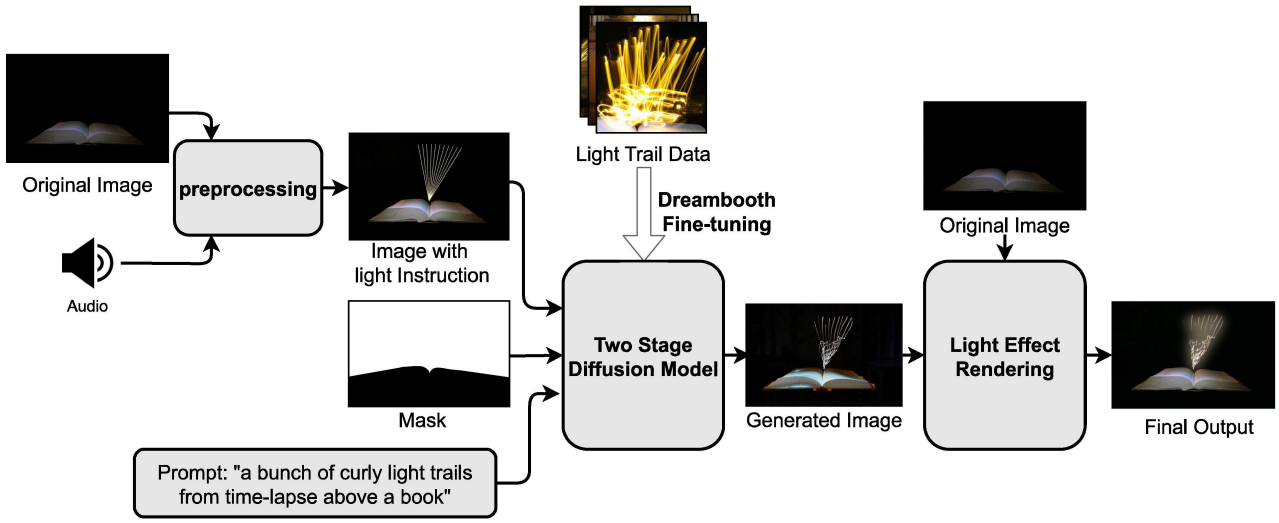


Figure 2. Process Flow for Real Light Trail Generation

more regimented approach using standard plotting functions to create regular, moving lines whose attributes are predefined yet responsive to the audio track. The visibility and intensity of these lines are governed by the 'brightness' feature, ensuring that their visual impact aligns with changes in the audio's volume and beat. As the audio's brightness increases, so does the prominence of these lines, making the visual effects not only more noticeable but also more expressive of the audio's dynamic properties.

3.2. Dreambooth Fine Tuning

In our research, DreamBooth fine-tuning is a critical step that adapts a general-purpose image synthesis model to specialize in generating light effects that are stylistically consistent with the chosen artist's work. This process leverages a subset of 67 curated images, each highlighting distinctive light trails from the artist's time-lapse photography, to train the model to replicate these unique visual elements.

The fine-tuning is performed using a pretrained Stable Diffusion model, tailored specifically to enhance its ability to generate dynamic light effects. We employ an adapted training regimen that include specific prompts related to light trails, adjustments to the learning rate and training duration to optimize performance, and a focus on high-resolution output to maintain quality in visual details.

This fine-tuning enables the model to produce images that aesthetically resonate with the original artworks light trail effect.

3.3. Utilization of the Fine-Tuned Model

After the fine-tuning phase, the adapted model is proficient at generating images that reflect the specific artistic style we targeted. Specifically, the model can now produce high-resolution images of dynamic light trails inspired by time-lapse photography, closely mimicking the desired artistic effects.

We utilize the fine-tuned model to generate new images based on predefined prompts that encapsulate the essence of the light trails characteristic of the artist's work. For instance, we might use a prompt such as "a bunch of curly light trails from time-lapse inside a train" to direct the model's output. This capability allows us to create images that are not only visually striking but also aligned with the specific aesthetic themes of the project.

The images generated post-training are used to evaluate the model's ability to replicate and innovate within the artistic parameters set by our project objectives. This step is crucial for demonstrating the practical application and effectiveness of our fine-tuning methodology.

3.4. Two Stage CustomPipeline

Our two-stage diffusion pipeline strategically divides the diffusion process into two parts to enhance the control and precision of light-effect generation in images. This model leverages the principles of Denoising Diffusion Probabilistic Models (DDPM) [2], Stochastic Differential Equation editing (SDEdit) [4], and inpainting methodologies [3] to refine the visual effects within specified areas of an image, focusing on enhancing the quality and integration of these effects seamlessly into the overall scene.

Stage One: During the initial stage, the model executes a typical diffusion sequence that generates preliminary visual content from a noise distribution towards a data distribution. This stage is mathematically described by the diffu-

sion equation:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sigma_t \epsilon$$

where X_t represents the image at time step t, ϵ_θ denotes the neural network predicting noise, and α_t and σ_t are parameters controlling the noise levels across the diffusion steps.

Stage Two: The second stage of the diffusion process employs a custom mask to selectively control the evolution of the image's latent representations. Upon reaching a predetermined step in the diffusion process, the model adjusts the latents according to the mask to focus enhancements on designated areas. This selective application ensures that only specified regions of the image are refined, maintaining the integrity of the overall composition while intensifying the effects where needed.

The adjustment of the latents at any step t after the mask is applied is described by the following equations:

$$x_{t-1}^m = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sigma_t \epsilon^m$$

$$x_{t-1} = x_{t-1}^m \odot m + x_{t'} \odot (1 - m)$$

Here, x_t represents the latent state of the image at time step t, x_{t-1}^m is the latent representation of the image in the masked areas after denoising at timestep $t-1$, $x_{t'}$ is the preserved latent state from the step when the mask was first applied, and m is the custom mask. The operator $\odot$ indicates element-wise multiplication, which allows the diffusion process to modify the image selectively as per the mask's guidance.

This methodological innovation allows for detailed refinement of visual effects within the framework of diffusion-based generative models. By applying the mask selectively, the model can enhance specific features or areas of the image without altering the entire scene, thus achieving a high degree of control over the visual outcome. This approach is particularly effective for applications requiring fine-tuned visual details in specific areas of an image while preserving the natural appearance of the rest of the scene.

3.5. Light Effect Rendering Algorithm

Inspired by the foundational work by Debevec et al. [1], we develop an algorithm to adjust the luminance of the background image based on the light effects from related foreground masked object masked. The goal is to maintain consistent illumination across the image by compensating for light differences caused by the object (light trail part).

The original image, model output image, and the final image used in the algorithm are shown in Fig. 3.

Initially, we compute the luminance contribution from the masked region of the model output image, which we term as the local scene luminance (LS_{obj}). We define the

(a) original image (b) model output image (c) final image

Figure 3. Images in Light Effect Rendering Algorithm

luminance weight vector $w = [0.3, 0.4, 0.3]$ to simplify our notation. The local scene luminance is computed as:

$$LS_{\text{obj}} = \sum(\text{model_output_image} \cdot \text{mask}) \cdot w \quad (1)$$

We then calculate the luminance for the unmasked region, representing the background unaffected by the object (essentially, $1 - \text{mask}$). We define this as the global scene luminance (LS_{noobj}):

$$LS_{\text{noobj}} = \sum(\text{model_output_image} \cdot (1 - \text{mask})) \cdot w \quad (2)$$

The final image luminance (LS_{final}) is computed by initially determining the total image luminance from the original image (LS_{total}) and adjusting it based on the calculated local and global scene luminance:

$$LS_{\text{total}} = \sum(\text{original_image}) \cdot w \quad (3)$$

$$LS_{\text{final}} = LS_{\text{total}} - (LS_{\text{obj}} - LS_{\text{noobj}}) \quad (4)$$

To adjust for any discrepancies in brightness due to the addition of the object, we scale the final luminance based on the average brightness comparison between the model output image and the original image:

$$\text{scaling_factor} = \frac{\text{avg_brightness}(\text{model_output_image})}{\text{avg_brightness}(\text{original_image})} \quad (5)$$

$$LS_{\text{scaled}} = LS_{\text{final}} \cdot \text{scaling_factor} \quad (6)$$

Finally, we apply the scaled luminance to get the color-corrected background:

$$\text{image}_{\text{final}} = \text{original_image} \cdot \frac{LS_{\text{scaled}}}{\text{original_luminance}} \quad (7)$$

This algorithm ensures that the light effect of the object in the mask reasonably contributes to the overall luminance consistency of the background, thus maintaining visual coherence throughout the image.

4. Conclusion

In this project, we have successfully extended the interactive possibilities of digital art through the synchronization of light effects with audio inputs, thereby enriching the visual and auditory experience. By integrating advanced image processing techniques with audio-driven dynamics, our system not only enhances the interactivity and aesthetic appeal of digital media but also paves the way for new forms of artistic expression.

The implementation of our image-to-image custom pipeline, especially the use of DreamBooth fine-tuning model, has demonstrated the capability to generate dynamic light trails that are responsive to audio signals. This approach has allowed us to produce visualizations that are not only vibrant and engaging but also maintain the integrity of the background, ensuring that the artistic vision is not compromised.

Our research contributes to the fields of computational creativity and digital arts by providing a framework that can be adapted to various artistic styles and preferences. The flexibility of our method, coupled with the precision of the light effects, offers artists and developers new tools to explore the interplay between sound and visual elements in real-time.

For future work, we aim to extend our model to handle real-time video generation, which would allow for dynamic audiovisual displays during live performances and interactive installations. This expansion will require the development of robust real-time processing capabilities to ensure seamless synchronization and high-quality video output.

We believe the fusion of cutting-edge technology with artistic expression as demonstrated in this project not only broadens the horizon of digital media but also invites further exploration into the untapped potential of audiovisual synchronization in art and entertainment.

References

[1] Paul Debevec, Tim Taylor, and Jitendra Malik. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *TOG*, volume 18, pages 189–198, 1998. 2, 4

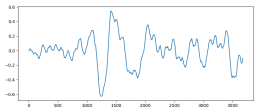
[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 1, 3

[3] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2201.09865*, 2022. 2, 3

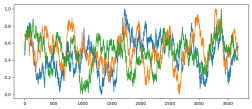
[4] Chenlin Meng, Jonathan Ho, Yang Song Chen, Pieter Abbeel, and Jiaming Song. Sdedit: Guided image synthesis and editing via stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021. 1, 2

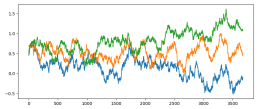
Appendix



(a) smoothed audio waveform



(b) noise is introduced audio waveform



(c) waveform with offset

(d) waveform mapped on the image

Figure 4. Waveform-based light trails mapping

(a) both custompipeline and dreambooth

(b) custompipeline and non dreambooth

(c) non custompipeline and dreambooth

(d) non custompipeline and non dreambooth

Figure 5. Comparison between applying custompipeline or not and applying fine tuning or not

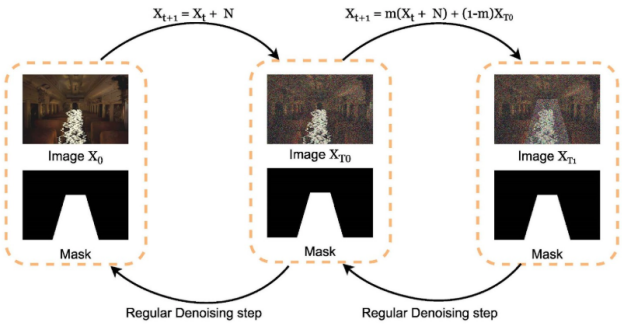
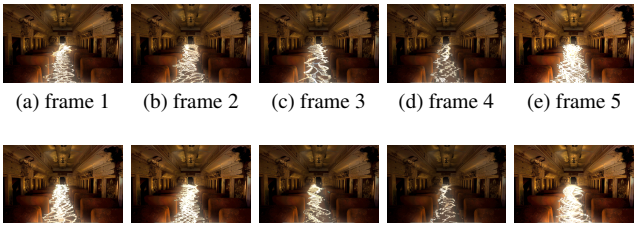


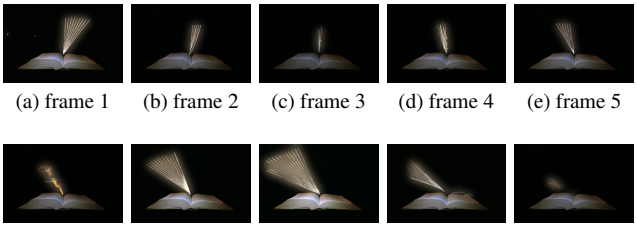
Figure 6. two-stage custompipeline visualization



(a) frame 1 (b) frame 2 (c) frame 3 (d) frame 4 (e) frame 5

(f) frame 6 (g) frame 7 (h) frame 8 (i) frame 9 (j) frame 10

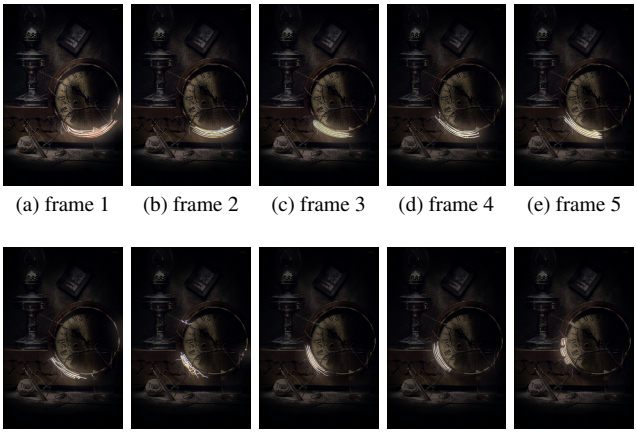
Figure 7. Frames of visualization: light inside a train



(a) frame 1 (b) frame 2 (c) frame 3 (d) frame 4 (e) frame 5

(f) frame 6 (g) frame 7 (h) frame 8 (i) frame 9 (j) frame 10

Figure 8. Frames of visualization: light above a train



(a) frame 1 (b) frame 2 (c) frame 3 (d) frame 4 (e) frame 5

(f) frame 6 (g) frame 7 (h) frame 8 (i) frame 9 (j) frame 10

Figure 9. Frames of visualization: light around a clock