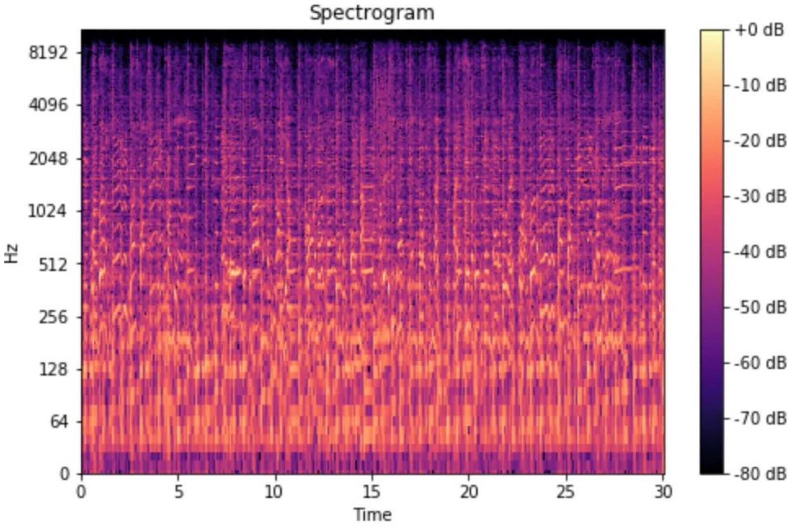
Images that Sound

Thanmaya Pattanashetty

Background: Spectrograms

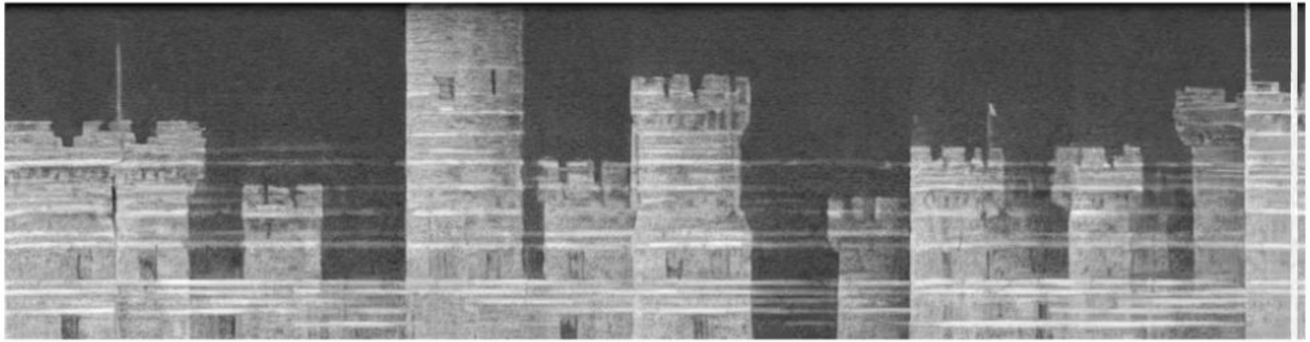
- **Spectrograms**: Visual representations of sound.
- Used in **audio machine learning** to depict sound features like **frequency and amplitude**.



Overview

- Goal: Create **images that sound**—spectrograms that are meaningful as both **images** and **audio**.
- Combines **text-to-image** and **text-to-spectrogram** diffusion models (Stable Diffusion & Auffusion).
- Opens new possibilities for **multimodal art** and **audio-visual learning**.

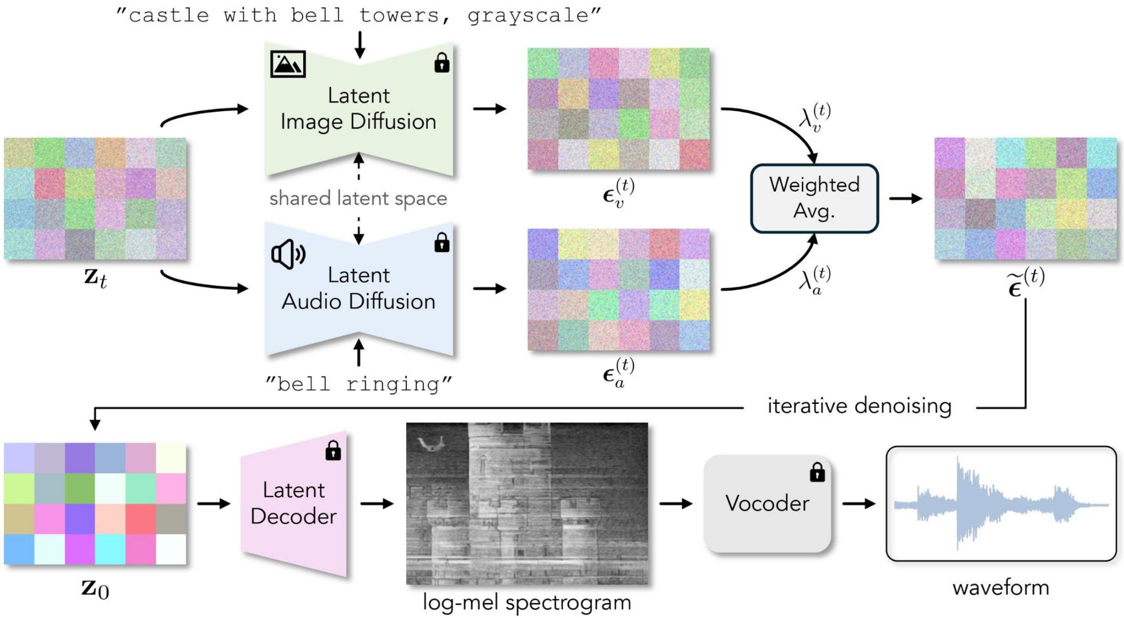
Image prompt: a painting of castle towers, grayscale



Audio prompt: bell ringing

Methods

- **Diffusion Models**: Iterative **denoising** to generate both images and sounds.
- **Multimodal Denoising**: Combines **audio** and **image diffusion models** using shared latent space.



Results

- **Metrics**: Evaluated using **CLIP** (image quality) and **CLAP** (audio quality).
- **Human Study**: Participants preferred the authors’ method for both **visual** and **audio** quality.
- Examples: E.g., a **castle** that looks like bell towers and sounds like **bells ringing**.

Limitations

- Cannot achieve **high-fidelity audio** and **high-quality visuals** at the same time all the time.
- Depends on well-crafted **prompts** for optimal results.
- Some visual and audio prompts do not work well together.

Conclusion

- Introduces **images that sound** using **diffusion models**.
- Potential in **art**, **cross-modal learning**, and **audio-visual applications**.
- Future improvements in **audio model quality** and multimodal interactions.

References

- Chen, Ziyang, Daniel Geng, and Andrew Owens. "Images That Sound: Composing Images and Sounds on a Single Canvas." *arXiv*, version 1, 20 May 2024, https://arxiv.org/pdf/2405.12221.
- Das, Saptarshi. "Understanding the Mel Spectrogram." *Medium*, 9 July 2020, https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.