# Integrating Multimodal Techniques with Latent Diffusion Models to Advance Multi-View Optical Illusion Generation

Wonseok Oh
University of Michigan
okong@umich.edu

Yeheng Zong
University of Michigan
yehengz@umich.edu

## Abstract

*We propose the approach to generate multi-view optical illusions using latent diffusion models enhanced by multimodal techniques. By incorporating sound and text inputs, our model expands the capability of traditional pixel-based diffusion processes, allowing for a richer, more dynamic generation of illusions. Central to our method is the integration of a denoising block designed to minimize latent-based artifacts, thus preserving the integrity and quality of the illusions. Our experimental results demonstrate the effectiveness of our approach, highlighted by the superior performance of our denoising techniques in improving the fidelity of generated illusions as assessed by CLIP scores. This research not only advances the technical framework for illusion generation but also opens new avenues for the application of multimodal inputs in the creative domain. Here is our implementation.*

## 1. Introduction

Geng [4] present a simple but effective method of generating multi-view optical illusions using diffusion models: given a noisy image $x_t$, the estimated noise $\epsilon_t^i$ is computed conditioned on text prompts $p_i$ and after applying transformation $v_i$. Geng *et al.* implemented their method on a pixel-based diffusion model to prevent the artifacts under rotation or flips when using a latent diffusion model. Based on their work, we expand sound as an additional modality that can be fused into the model. We also explore the possibility of replacing the pixel-based diffusion model with a latent one. Our potential contributions are:

- Generate multi-view optical illusions using both sound and text as the input prompts.
- Mitigate latent-based artifacts with a denoising block with a designed denoising schedule.

## 2. Related work

We aimed to utilize multimodality, mainly focusing on text, sound, and image modalities. Based on these, we have designed and implemented a diffusion model [6] to generate practical outputs for specific applications. Throughout this process, the resultant product was a latent diffusion model. We refined this model by incorporating a denoising step, enabling us to produce the final outputs. This approach underscores our commitment to leveraging advanced multimodal technologies to enhance application-specific performance.

### 2.1. Diffusion Models

The diffusion model aims to gradually recover ground truth signal $x_0 \sim p(x_0)$ added random noise $\epsilon_t \sim N(0, I)$ to desired images. To be more specific, the forward diffusion process $p(x_T|x_0)$ utilizes a Markov Chain that gradually mitigates $x_0$ to $x_T$ with random Gaussian noise.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

where $\beta_t \in (0, 1)$is the noise scale. Following the noise scheduler, $\beta_t$ increases as the timestep grows, and finally, ground truth images are completely covered with noise.

$$\begin{aligned} q(x_t|x_0) &= \prod_{t=1}^{T} \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}, \beta_t I) \\ &= \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0, \sqrt{1-\bar{\alpha}_t}I) \end{aligned} \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha_t} = \prod_{i=1}^{t} \alpha_i$. The diffusion model $\epsilon_\theta(x_t, t)$ is training to estimate $\epsilon_t$ from $x_t$, by gradually remove noise from the $x_t$. The backward process, commonly known as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

### 2.2. Illusions and Modality

Recent advancements in diffusion models have enabled artists and researchers to create complex illusions. For instance, Mr. Ugleh repurposed a QR code-generating model

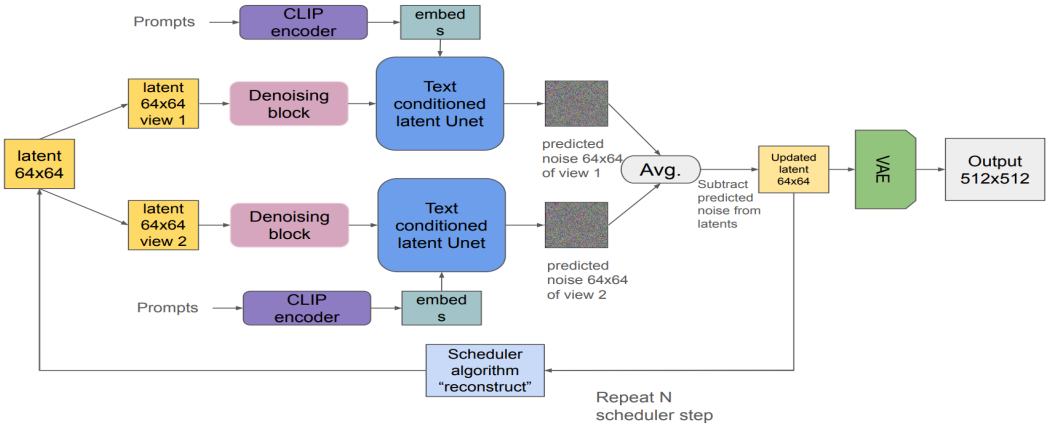


Figure 1. Overview of the entire model of the main model architecture. This obtains two inputs to generate the image with illusions.

to create images with a global structure matching specific templates, whereas Burgert et al. [2] used score distillation sampling [10, 8] to create multi-view images that change appearance based on the viewer's angle, albeit with lower quality and longer processing times. Geng introduced the algorithm with simultaneously denoising multiple views of an image. This work uses the pixel-based diffusion model to generate the images. Our model builds on the latent diffusion framework, adding a text prompt with an audio input. Unlike Geng's approach, which processes multiple image views simultaneously, our model enriches the illusion generation by integrating auditory stimuli, offering new possibilities for applications of modalities in the diffusion model. This approach enhances both the text similarity and dynamicity of the generated illusions. Here, we used a pre-trained VGG-based encoder to translate sound data into latent code.

## 3. Methods

Our goal is to make a latent diffusion model with multi-modality. To expand the modality, we add sound as another modality. To mitigate the artifacts of multi-view illusion generation when using the latent diffusion model, we further introduce a denoising block with the designed scheduler.

### 3.1. Conditioned Diffusion Model

Fig 1 is the main architecture for image processing that integrates a latent diffusion model. In our framework, an initial decoding block precedes the diffusion process. Images are first passed through a decoder, where they undergo a specified transformation $v$. The transformed images are then input into a latent diffusion model for further processing.

Following the diffusion model, a post-processing step involves applying the inverse transformation $v^{-1}$ to revert the image back to its original form. The image is then encoded to predict the noise component, completing the cycle.

The architecture is designed for flexibility; the decoder and encoder surrounding the diffusion model need not be complex neural networks. Alternatives include utilizing Discrete Fourier, cosine, and wavelet transforms as the decoders, with their corresponding inverse operations serving as the encoders.

Upon further consideration, it was decided to simplify the initial dual decoder-encoder design to avoid complications during the analysis and experimentation stages. The replacement is a singular image processing block that directly manipulates the latent. Options for this block include a Gaussian blurring filter to smooth the image, a set of filters that induce smooth transitions in the latent space, and a component that suppresses high-frequency elements by zeroing out certain coefficients post-discrete Fourier Transform.

### 3.2. Sound data as an input

As illustrated in Fig 2, we designed the model with the data types of the two prompts originally used as inputs for our main model to facilitate multi-modality. This modification involved receiving one input in the form of a text prompt and another as audio data. Similar to the process for text, the input audio is processed through a VGG-based encoder [3], which we refer to as the New Imagebind-based Encoder [5]. This step allows us to place the data into an embedded space. Following this, the data undergoes a denoising step via a UNet architecture. This approach effectively integrates diverse input modalities, enhancing the model's capability to handle and refine multimodal data inputs.

### 3.3. Fourier Denoising

To mitigate the latent-based artifacts, we use a Fourier denoising block. The Fourier denoising block cuts off percentages of the Discrete Fourier Transform(DFT) coefficients by their magnitude and then does reconstruc-

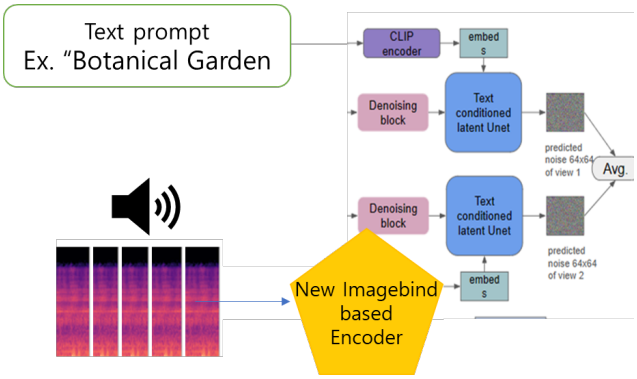


Figure 2. Designed model with two inputs which are in the different domains and different modalities

tion using the deducted DFT coefficients. Previous research [7] states that the latent bases evolve from low- to high-frequency components, and hence, we make the strength of denoising (percentage of DFT coefficients being cut) decrease as the step goes further to stay consistent with this finding. In our implementation, the Fourier denoising operation is performed on each channel of the latent starting at step 200 (with a total of 500 steps) with an interval of 20 steps. The cut-off percentages of DFT coefficients from step 200 to 300 is 50%, from step 300 to 400 is 40%, and from step 400 to 500 is 30%.

### 3.4. Wavelet Denoising

Cutting off the DFT coefficient matches the fact the latent space is closely related to frequency. However, we disregard the complex information in DFT and only use the real information. To prevent missing information, we further investigate a Wavelet denoising block. Inside the Wavelet denoising block, a soft-thresholding operator is applied to every Discrete Wavelet Transform(DWT) coefficient. Mathematically, for a DWT coefficient $c$, $\text{soft}(c;\lambda) = \text{sgn}(c) \cdot \max(0, |c| - \lambda)$ given the value after soft-thresholding with threshold $\lambda$. In our implementation, the Wavelet denoising operation is performed on each channel of the latent starting at step 200 (with a total of 500 steps) with an interval of 20 steps. We use Daubechies Wavelet at a level of 4. The threshold of the soft-thresholding operator from step 200 to 300 is the 94% percentile of the magnitude of DWT coefficients, from step 300 to 400 is the 88% percentile of the magnitude of DWT coefficients, and from step 400 to 500 is the 82% percentile of the magnitude of DWT coefficients.

### 3.5. TV Regularization Denoising

In both Fourier and Wavelet denoising blocks, some percentage of coefficients are set to zero, which may result in loss of information. To prevent this problem, we choose Total Variance(TV) regularization denoising as the third option of the denoising block. The total variance of a 2D image $x$ is defined as:

$$\text{TV}(x) = \sum_{i,j} \sqrt{|x_{i+1,j} - x_{i,j}|^2} + \sqrt{|x_{i,j+1} - x_{i,j}|^2}$$

The TV regularization denoising [1] is indeed solving the optimization problem $\hat{x} = \arg\min_{x \in \mathbb{R}^{\#}} \|y - x\|_2^2 + \lambda \cdot \text{TV}(x)$, where $y$ is the observed data, $x$ is initialized as a copy of $y$, and $\lambda$ is a positive regularization constant. In our implementation, we solve the optimization through Proximal Gradient Descent. The step size is $3 \times 10^{-5}$, and the number of iterations is 300. The TV regularization denoising is performed on each channel of the latent starting at step 200 (with a total of 500 steps) with an interval of 20 steps. The value of regularization term $\lambda$ decreases as steps go further: $\lambda = 0.18$ from step 200 to 300 , $\lambda = 0.09$ from step 300 to 400, and $\lambda = 0.045$ from step 400 to 500.

| CLIP Score | view1 | view2 |
|---|---|---|
| Raw | 0.7317 | 0.6527 |
| Fourier | 0.8098 | 0.7511 |
| Wavelet | 0.7415 | 0.6974 |
| TV Reg | 0.7447 | 0.6929 |

Table 1. **Quantitative Results**. The table records the CLIP score of the generation from the raw model and from models with different denoising blocks. The CLIP score of generations from models with all three denoising blocks performs better than the raw generation.

## 4. Experiment

### 4.1. Quantitative Evaluation

We use the CLIP Score [9] to evaluate our text-conditioned generation with a denoising block. The CLIP score between an image $I$ and text $T$ is calculated based on their cosine similarity in a common feature space. The mathematical expression for the CLIP score is:

$$\text{CLIP Score}(I, T) = \frac{\langle f(I), g(T) \rangle}{\|f(I)\| \|g(T)\|}$$

$f(\cdot)$ is the CLIP image encoder and $g(\cdot)$ is the CLIP text encoder. Hence, we expected that the CLIP score would stay at approximately the same level with and without the denoising block. We summarized our CLIP score evaluation in Table 1. Surprisingly, we find out that models with denoising block has higher CLIP Score

### 4.2. Qualitative Evaluation

As illustrated in Fig 3, the denoising block helps mitigate the artifact and smooth the image to have better visual quality. Fig 4 shows a interesting phenomenon in Fourier

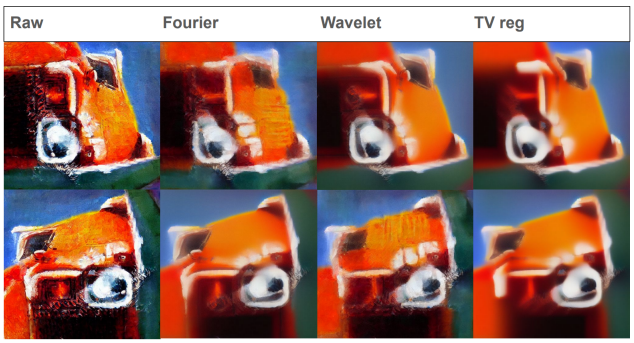


Figure 3. Generation with text prompts "a painting of truck" (first row) and "a painting of red panda" (second row)

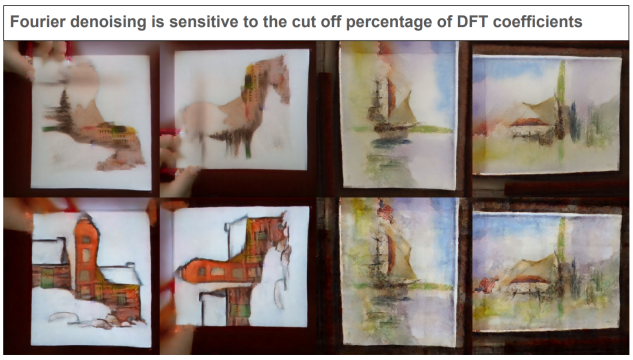


Figure 4. Fourier denoising with different cut-off percentage of DFT coefficients may change the content in generation. The difference between first row and second row in cut-off percentage is $10\%$

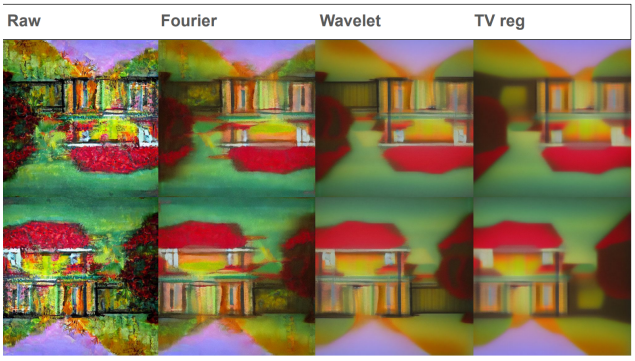


Figure 5. Generation with text prompts "an oil painting of botanical garden" (first row) and "an oil painting of house" (second row)

denoising: Fourier denoising with different percentage of cut-off DFT coefficients may even change the content in generation. Fig 5 shows a case that the denoising block do not have a positive effect on the generation. The detailed and rich texture of the botanical is deducted by the denoising block.

| CLIP Score | view1 | view2 |
|---|---|---|
| woF | 0.7258 | 7117 |
| woW | **0.7516** | **0.7190** |
| woTV | 0.7151 | **0.7029** |

Table 2. **Ablation study**. The table records the CLIP Score of each denoising block without scheduling. We notice that the scheduling of changing the scale of denoising (or strength of the regularization) will not always lead to a higher CLIP Score. The Wavelet denoising without scheduling of changing denoising scale is better than Wavelet denoising with scheduled denoising scale under CLIP Score evaluation

### 4.3. Denoising Scheduler Evaluation

To investigate the effectiveness of changing the denoising scale as the step goes further, we evaluate the CLIP Score of generations without changing the denoising scale. For all of the three denoising block, denoising starts at step $200$ (with total $500$ steps) with interval of $20$ steps. For Fourier denoising, the cut-off percentage of the DFT coefficient stays at $40\%$; for Wavelet denoising, the threshold is always the $88\%$ percentile of the magnitude of DWT coefficients; and for TV regularization denoising, the regularization term $\lambda$ stay at $0.045$. We find that only Fourier denoising always benefits from the scheduled denoising. The result may reflect two pieces of information. First, the change of threshold value in Wavelet denoising and the change of regularization strength in TV denoising can be tuned to be better. Second, it would be better to make the change of threshold value and regularization strength become data-adaptive.

## 5. Conclusion and Limitation

In this project, we expand sound as an additional modality to generate multi-view illusions using latent diffusion models. With the help of a denoising block, we mitigate the latent-based artifacts in the generation. However, there are three main limitations of this simple method. First, we do not have enough view transformation (only rotation by $90$ and $180$ degrees for now) and hence the generation of a wide range of text prompts pair is poor, i.e., the generation conditioned on "a painting of house plant" and "a painting of Albert Einstein". Second, we naively draw equivalence between artifacts, noise and rich texture. The botanical garden generation is a typical example reflect this problematic assumption. Third, the latent-based artifacts may also be data-based artifacts. Since we are using a pre-trained latent diffusion model, the lack of training data of some specific type may also caused the artifact. To overcome this limitation, a more comprehensive research is required on the cause of these artifact, and probably adding image prior can help with fix it.

**Acknowledgements** We appreciate the discussion with Professor Andrew Owens and Professor Liyue Shen. They gave us valuable suggestions. We also appreciate Xiaofeng Dai for his close collaboration with us.

## References

[1] Emeric Boigné, Dilworth Y Parkinson, and Matthias Ihme. Towards data-informed motion artifact reduction in quantitative ct using piecewise linear interpolation. *IEEE Transactions on Computational Imaging*, 8:917–932, 2022.

[2] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael S Ryoo. Diffusion illusions: Hiding images in plain sight. *arXiv preprint arXiv:2312.03817*, 2023.

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[4] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. *arXiv preprint arXiv:2311.17919*, 2023.

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[7] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[9] 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.

[10] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.