Enhancing Multi-View Optical Illusion Generation with Latent Diffusion Models and Traditional Image Processing

Wonseok Oh Xiaofeng Dai Xingjian Jiang Yeheng Zong

EECS556 Image Processing

April 26, 2024

1 Honor Code

Our team will give attribution for any figures used in our documents and will cite all code sources. Wonseok Oh and Yeheng Zong use the same main reference paper Visual Anagrams by Geng *et al.* [1] and also explore on latent diffusion model in their project in EECS 542. The contents and focuses of these two project are different. In the EECS 542 project of Wonseok Oh and Yeheng Zong, they incorporate sound and text inputs and hence expand the capability of traditional pixel-based diffusion processes, allowing for a richer, more dynamic generation of illusions. In this project for EECS 556, the focus is on blending in image processing techniques (motion illusion and TV regularization denoising) to reduce artifacts. Hence, all the efforts and contributions in this report belongs to EECS 556.

2 Introduction

Images have consistently held significance in human existence, as vision likely stands as the most crucial sense for humans. The domain of image processing represents an interdisciplinary convergence of computer science and digital signal processing aimed at the meticulous analysis and refinement of visual data to extract pertinent insights or enhance perceptual accuracy.[2]. However, the challenges posed by poor quality of images and insufficient image data significantly hinder the effectiveness of various algorithms and systems designed for tasks including object recognition, image classification, and enhancement[2]. Consequently, many researchers focus on devising robust methods for data augmentation, image enhancement, and the development of algorithms capable of learning

from limited or compromised visual information. Conventional methodologies for image enhancement predominantly operate within the spatial and frequency domains. While histogram equalization[3] stands as a prevalent technique in this domain, its global adjustment approach proves inadequate in effectively augmenting local contrast. As a result, Wang and Pan[4] introduced a novel approach that partitions the image into active, inactive, and general areas based on pre-established gradients, thus facilitating the targeted selection of local regions within the image. Moreover, The emergence of deep learning architectures, including Generative Adversarial Networks (GANs)[5], Variational Autoencoders (VAEs)[6], and Diffusion Models[7], has significantly enhanced the capability to generate high-quality images.

However, these methods are demonstrated to be constrained by human visual perception. Given that human visual perception of images is limited to a single viewpoint, details are obscured or less noticeable in traditional 2D images, posing obstacles to comprehending the entire spectrum of features within the image[8]. The paper[1] addresses the complexity of fabricating multi-view optical illusions through the utilization of diffusion models and text conditioning. By employing diffusion models to cleanse images from various viewpoints and integrating textual prompts as conditioning factors, the approach generates illusions that metamorphose under transformations like rotations, flips, and jigsaw rearrangements[1]. This methodology offers extensive flexibility in crafting illusions that alter their appearance under diverse transformations, facilitating the synthesis of a broad spectrum of dynamic visual effects that push the boundaries of human perception. Furthermore, the diffusion models in [1] contribute to the optimization of the quality of generated illusions. By analyzing the conditions under which transformations are supported and making design decisions to enhance the illusion generation process, diffusion models help ensure the efficacy and flexibility of the method. Therefore, our project aims at exploring and expanding the capability of deep learning models in[1] to synthesize image transformations that improve the diversity of the original dataset and enhance the quality of the resulting images. By employing diffusion models and textual prompts, our approach seeks to generate illusions that change with image transformations offering new elements for the crafting of illusions with multiple viewpoints and synthesizing new images. Our key research question focuses on how deep learning, specifically latent diffusion models, combined with traditional image processing tools taught in EECS 556, can be optimized to create high-quality, multi-view optical illusions that transcend conventional visual perceptions. The potential contributions of our project are as follows:

- Generate multi-view optical illusions to improve the perception of details being obscured or less noticeable in traditional 2D images.
- Combined diffusion model and traditional image processing tools to mitigate artifacts
- We provide quantitative and qualitative results to demonstrate the flexi-

bility of our method.

We first reproduced the results shown in the Visual Anagrams model to generate images that have 90° rotation multi-view optical illusions with the published code.[1] and the prompts they tested as reference of performance. We then experimented to combine the latent diffusion model with traditional image processing tools (Motion illusion [9] and denoising method in this work)

3 Related Work

3.1 Diffusion Models

The diffusion model aims to gradually recover ground truth signal $x_0 p(x_0)$ added random noise $\epsilon_t N(0, I)$ 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, β_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 ϵ_t from x_t, by gradually remove noise from the x_t. The backward process, commonly known as

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

3.2 Progress of the related works

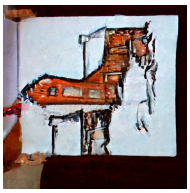
3.2.1 Motion Illusion

To overcome the limitation in the original model that the artifacts of latent representations appear under transformations like rotation and flip, we will implement the motion illusion [9] as a constraint to eliminate the artifact from the latent domain. The method uses a quadrature pair of oriented filters to vary the local phase, giving the sensation of motion. We will first discuss the content together and develop the module to implement motion illusion in our images. Then we plan to test and fine tune the parameters of the filters in the module to let motion illusion matches different transformations. We then want to test if this method will give us the correct response of both the location and orientation of latent features from the images. If so, we will compare the performance of our method with their pixel-based diffusion model in the aspect of how the results correspond to the text prompt.

(a) An oil painting of a red panda

(b) An oil painting of kitchenware

Figure 1: An Example of Vertical Flip Visual Anagram



(a) A horse

(b) A snowy mountain village

Figure 2: An Example of Rotation Visual Anagram in Cartoon Drawing Style

3.2.2 Multi-view Denoisng with Latent Diffusion Model

At this first stage, we implement the protocol of multi-view denoising with latent diffusion model. Our implementation is based on Stable Diffusion tutorial and Visual Anagrams. In our implementation, we presented two strategies. The first one alternatively denoised two views, while the second one denoised two views in one iteration and then averaged the predicted noise. For simplicity, we have only implemented two possible view transformations now, which are the 90-degree rotation and vertical flip, but we will expand the list of available views in the future.
We present two figures (Fig 1 and Fig 2) of our preliminary generation results. We reproduce the latent-based artifact mentioned in [1]. When using latent diffusion models, we see artifacts that force straight lines to be thatched under flip and rotation. Encoding the image to latent involves a convolution operation, and hence, manipulating the location of the latent representation does not change the orientation of the pixel blocks after the de-convolution operation. We aim to mitigate these artifacts by blending in image processing techniques, and we will discuss the challenges and further plan in the later section.

3.2.3 Preliminary evaluation

We have embarked on the preliminary evaluation of our model, leveraging widely recognized metrics in the domain of diffusion models. Specifically, we employ the CLIP Score and PIQE Score. Our evaluation process begins with the application of these metrics to two distinct scenarios, each with a unique prompt designed to test the model's capability in generating images based on specific textual

descriptions. The prompts and their corresponding image-generation tasks are as follows:

1. **Scenario One:**
    - *Prompt 1:* "A watercolor of a ship"
    - *Prompt 2:* "A watercolor of a great view"

Figure 3: Images of Scenario One

2. **Scenario Two:**
    - *Prompt 1:* "A cartoon drawing of a horse"
    - *Prompt 2:* "A cartoon drawing of a snowy mountain village"

Figure 4: Images of Scenario two

The generated images for each prompt were then subjected to evaluation using our chosen metrics. The preliminary results are summarized in the table below:

Table 1: Preliminary Evaluation Results of Generated Raw Images

Prompt	**CLIPScore**
A watercolor of a ship	0.8725
A watercolor of a great view	0.7280
A cartoon drawing of a horse	0.7540
A cartoon drawing of a snowy mountain village	0.6033

Using the following method, we created the final evaluation result using various denoising ideas with motion illusion. The results are as follows.

4 Quantitative performance prediction

4.1 CLIP Score

We use the CLIP Score [10] 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.

4.2 PIQE Score

The Perception-Based Image Quality Evaluator (PIQE) is a no-reference tool that is used to assess the perceptual quality of images without comparing them to a reference image. It works by calculating the Mean Subtracted Contrast Normalized (MSCN) coefficients for each pixel to identify textual and noise variations. Here, we use this tool in MATLAB to make the quantitative results for the generated images because it is a metric that checks the quality of the images without any image to compare due to the characteristics of PIQE. This is meaningful in checking the quality of the image itself that we generated.

5 Methods

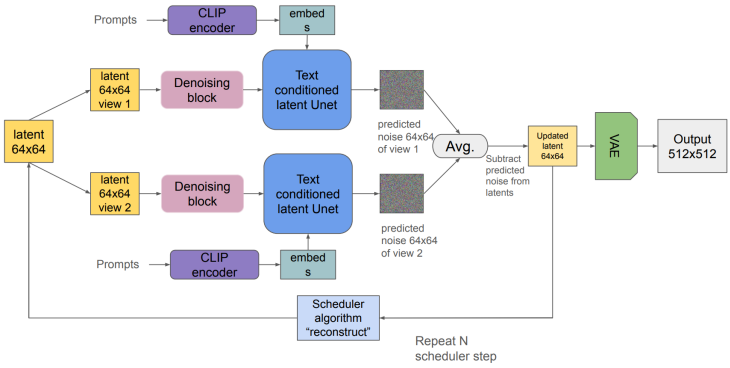


Figure 5: Overview of the entire model of the main model architecture. We adapted the simple but efficient method introduced by Geng *et al.* [1] with blending in an image processing that aim to mitigate the latent-based artifacts.

To mitigate the artifacts of multi-view illusion generation when using the latent diffusion model, we further introduce an image processing block with the designed scheduler.

5.1 Conditioned Diffusion Model

Fig 5 is the main architecture for image processing that integrates a latent diffusion model. In our framework, an initial image processing block precedes the diffusion process. Options for this image processing block include motion illusions, Fourier and Wavelet denoising, and Total Variance Regularization denoising, and we will discuss each of them in the later subsections. The latent is first passed through this block, where it undergoes specified processes that aim to mitigate the artifacts in the final output. The processed latent is then transformed into two views, and each of them will be inputted into a latent diffusion model for the diffusion process, which is conditioned on different text prompts. The latent is updated by subtracting the average of predicted noises from two views. We repeated this step for N times, and the final output that represents a multi-view visual anagram can be obtained by passing the latent in the last step into a Variational Auto-Encoder.

5.2 Motion Illusion

Here we applied the method developed by Freeman et al to display patterns that appear to move continuously without changing their positions.[9] The sensation of motion is achieved by a quadrature pair of oriented filters to vary local phase over time, which are identical except shifted in phase from each other by 90 degrees and are related by the Hilbert transformation. Here, we used the second derivative of a Gaussian, G_2, and its Hilbert transform, H_2. To introduce the variation in time, we construct the sequence of phase-shifted filters as shown below:

$$F(t) = cos(\omega t)G_2 + sin(\omega t)H_2$$

where F is the phase-shifted filter, ω is the rate of shift, and t is time. To change the orientation of F, we synthesize G_2 from a linear combination of basis filters:

$$G_2^\theta = k_1(\theta)G_{2a}(x,y) + k_2(\theta)G_{2b}(x,y) + k_3(\theta)G_{2c}(x,y)$$

with $k_i(\theta)$ stands for interpolation functions and $G_{2q,2b,2c}(x,y)$ represents basis functions for $G_2^\theta(x,y)$. Respectively, we can build the H_2^θ with the same method. The visualization of the basis is shown in Fig 6:

To make an image $I(x,y)$ appear to move in a direction, $\theta(x,y)$, at every point (x,y) in the image, we perform the calculation below to get even and odd phase images:

$$\begin{aligned} E(x,y) &= I(x,y) \otimes G_2^\theta(x,y) \\ O(x,y) &= I(x,y) \otimes H_2^\theta(x,y) \\ D(x,y,t) &= cons(\omega t)\, E(x,y) + sin(\omega t)\, O(x,y) \end{aligned}$$

This enables us to calculate image sequence $D(x,y,t)$ by taking the weighted sum of even and odd phases, where ω is the temporal frequency of the motion. We applied this method to a sample circle disk at different angles to show the

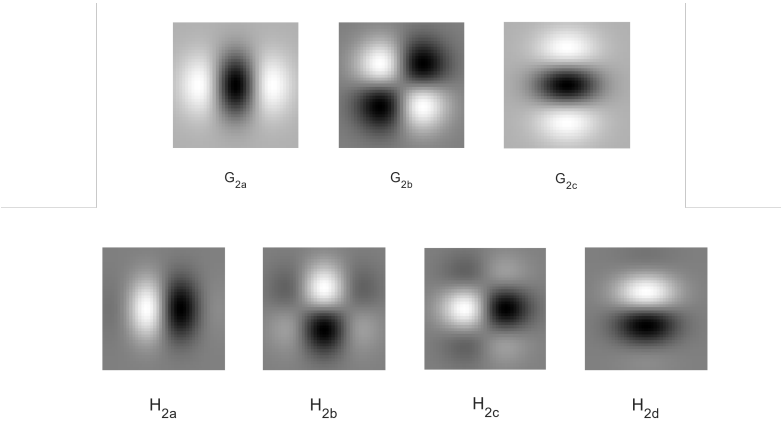


Figure 6: X-Y separable basis filter for G_2 and H_2

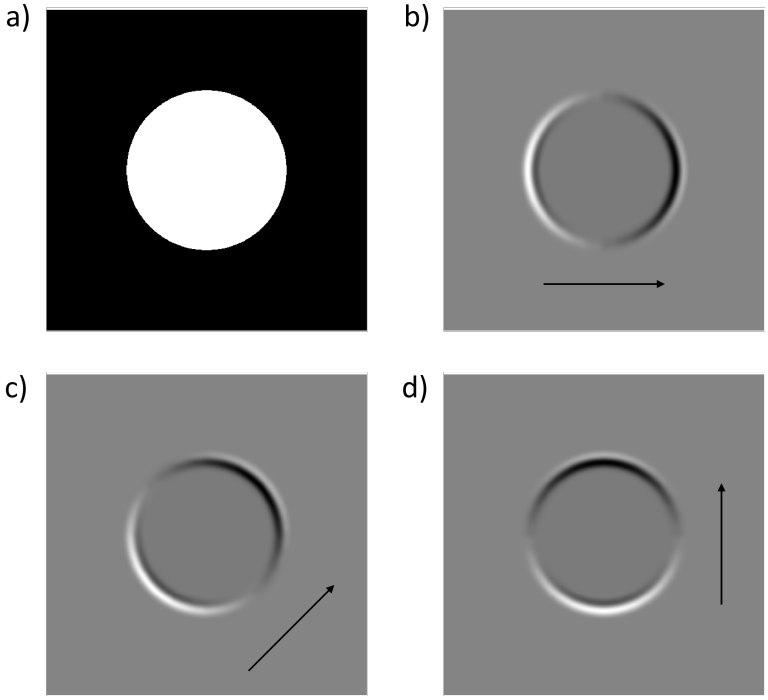


Figure 7: Example of phase-shifted filters applied to a circular disk. a) image of circular disk; b) $\theta = 0°$; c) $\theta = 45°$; b) $\theta = 90°$

performance of this phase-shifted filter on basis element patterns. (Fig 7) We also tested with real images to show the vision illusion of motion as a temporal

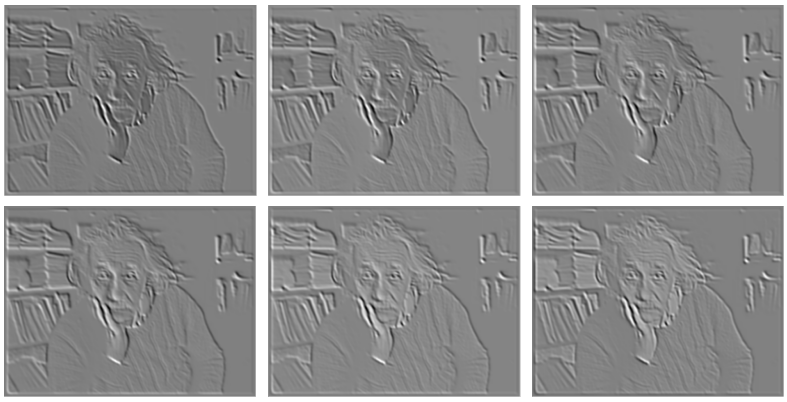


Figure 8: Example of phase-shifted filters applied to an image of Einstein. This generates the perception of rightward motion with images remaining stationary

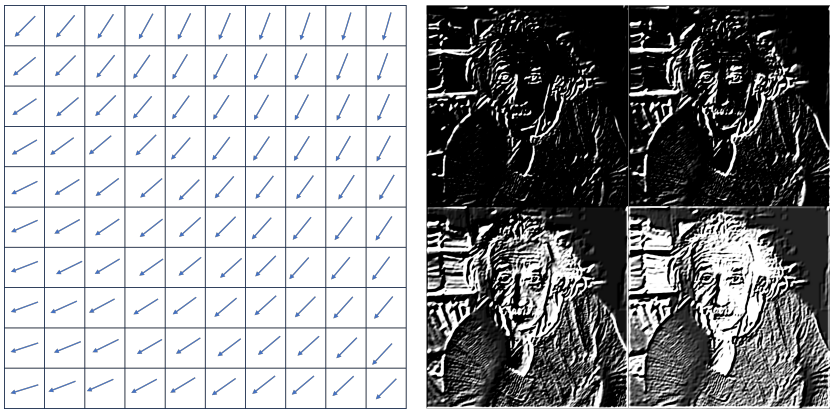


Figure 9: Left. Segments of image regions and motion orientation of each region from the rotation operation. Right. Rotation motion illusion is generated by a phase-shifted filter according to the orientations shown on the left. A continuous temporal sequence can be seen in our GitHub link

sequence (Fig 8). To make this applicable to the task of mitigating artifacts in the multi-view illusion images generated by the latent-based diffusion model, we calculated the orientation of each sub-region, which generates a rotation illusion for the raw image. And we applied our phase-shifted filters with the orientation calculated as input to the image and assembled the fragments to get the final image. (Fig 9)

We observed that the latent space has a noticeable similarity to the phase image of the final output. And the artifacts mostly show up at the edges. So this inspired us to apply the motion illusion filter to a single channel of latent space. The major multi-view illusion we are experimenting with in this project

is the 90° rotation. So we implemented the rotational version of motion illusion filter (Fig 9) to the latent space, our goal is to reinforce the rotation by this filter so that the artifacts can be smoothed in the final output. We added this filter every 50 steps after step 300, considering the edge features are not very obvious in the early steps.

5.3 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 reconstruction using the deducted DFT coefficients. Previous research [**?**] 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%.

5.4 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 λ. 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.

5.5 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 [**?**] is indeed solving the optimization problem $\hat{x} = \arg\min_{x \in R^2} \|y-x\|_2^2 + \lambda \cdot \mathrm{TV}(x)$, where y is the observed data, x is initialized as a copy of y, and λ is a positive regularization constant. In our implementation, we solve the optimization through Proximal Gradient Descent. The step size is 3×10^{-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 λ 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.

6 Evaluation

6.1 Quantitative Evaluation

6.1.1 CLIP Score Evaluation

Here, we use the CLIP Score [10] 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 2. Surprisingly, we found out that models with denoising blocks have a higher CLIP Score.

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 2: CLIPscore 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.

6.1.2 PIQE score Evaluation

The Perception-Based Image Quality Evaluator (PIQE) is a no-reference tool that is used to assess the perceptual quality of images without comparing them

to a reference image. It works by calculating the Mean Subtracted Contrast Normalized (MSCN) coefficients for each pixel to identify textual and noise variations. Here, we use this tool in MATLAB to make the quantitative results for the generated images because it is a metric that checks the quality of the images without any image to compare due to the characteristics of PIQE. This is meaningful in checking the quality of the image itself that we generated. The result of TV regularization gives the best quantitative result. The Fourier Denoising and Wavelet Denoising decrease the image quality, which can be reasonably explained by the fact that setting coefficients to zero results in loss of information.

PIQE	view1	view2
Raw	19.8757	19.3366
Fourier	23.4951	22.9707
Wavelet	20.9396	21.0955
TV Reg	**19.3015**	**19.2084**

Table 3: PIQE score Results. PIQE score gives us individual image quality benchmarks. In the official directory of MATLAB, there is a reference for the point. If the point is lower than 20, this refers that the image is in excellent quality. The result of TV regularization gives the best quantitative result.

6.2 Qualitative Evaluation

As illustrated in Fig 10, the denoising block helps mitigate the artifact and smooth the image to have better visual quality. Fig 12 shows an interesting phenomenon in Fourier denoising: Fourier denoising with different percentages of cut-off DFT coefficients may even change the content in a generation. Fig 13 shows a case in which the denoising block does not have a positive effect on the generation. The detailed and rich texture of the botanical is deducted by the denoising block. Fig 11 shows an extra generation result.

6.3 Denoising Scheduler Evaluation

To investigate the effectiveness of changing the denoising scale as the step goes further, we evaluate the CLIP Score and the PIQE 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 λ stay at 0.045. For CLIP Score, We find that only Fourier denoising always benefits from the scheduled denoising. For PIQE score, all of them are higher and hence imply the decrease in image quality. The result may reflect two pieces

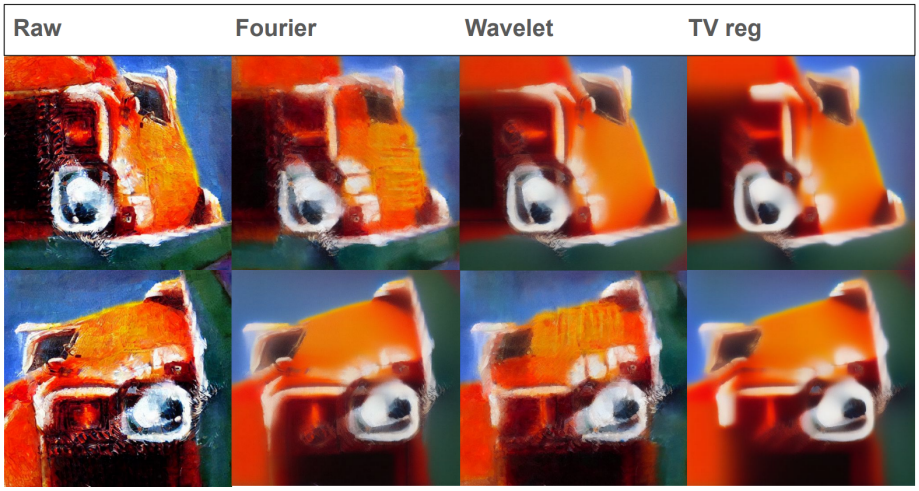


Figure 10: Generation with text prompts "a painting of truck" (first row) and "a painting of red panda" (second row)

Figure 11: Generation with text prompts "a watercolor painting of a ship" (first row) and "a watercolor painting of a village in the mountains" (second row)

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.

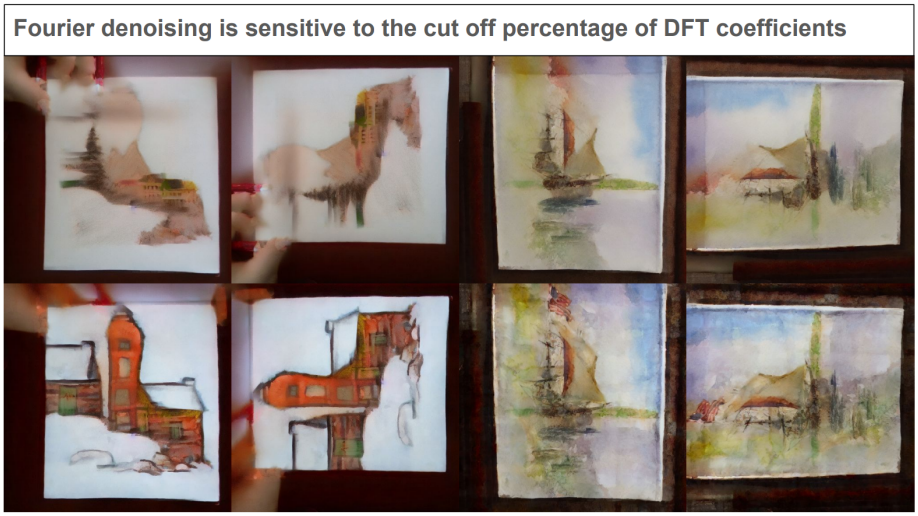


Figure 12: Fourier denoising with the different cut-off percentages of DFT coefficients may change the content in a generation. The difference between first row and second row in cut-off percentage is 10%

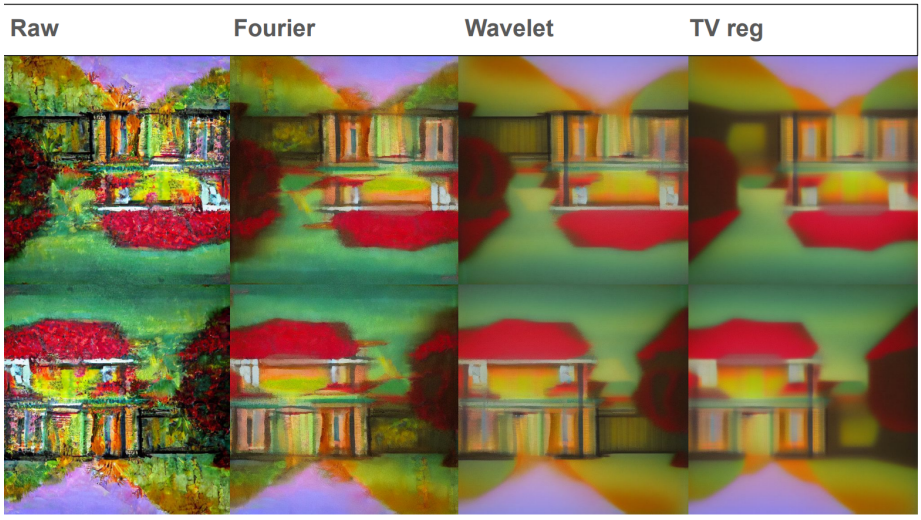


Figure 13: Generation with text prompts "an oil painting of botanical garden" (first row) and "an oil painting of house" (second row)

7 Conclusion and Limitation

In this project, with the help of an image processing block, we mitigate the latent-based artifacts in the generation. However, there are four main limita-

CLIP Score	view1	view2
w/o F	0.7258	7117
w/o W	**0.7516**	**0.7190**
w/o TV	0.7151	**0.7029**

Table 4: 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

PIQE Score	view1	view2
w/o F	22.7629	22.9361
w/o W	17.2212	17.0809
w/o TV	16.4636	16.1434

Table 5: Ablation study. The table records the PIQE Score of each denoising block without scheduling. We notice that the scheduling of changing the scale of denoising (or strength of the regularization) will lead to a higher PIQE score and hence a lower image quality.

tions of this simple method.

7.1 Missing Magnitude Information

The motion illusion didn't work very well in this task. As we can see in Fig 14, when applying this filter to latent space, we lose the magnitude information, which influenced the decoding and thus messed up the final output. To overcome this issue, predicting the magnitude of information from other channels and previous runs can be promising.

7.2 Inadequate Transformation Views

First, we do not have enough view transformation. We have only implemented rotation for 90 and 180 degrees for now. However, these simple transformations remain the majority of patterns, textures, and contents in the image before and after the transformation. In other words, the content of two views must be relatively closely related to each other (for example, both horse and house are brown), 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."

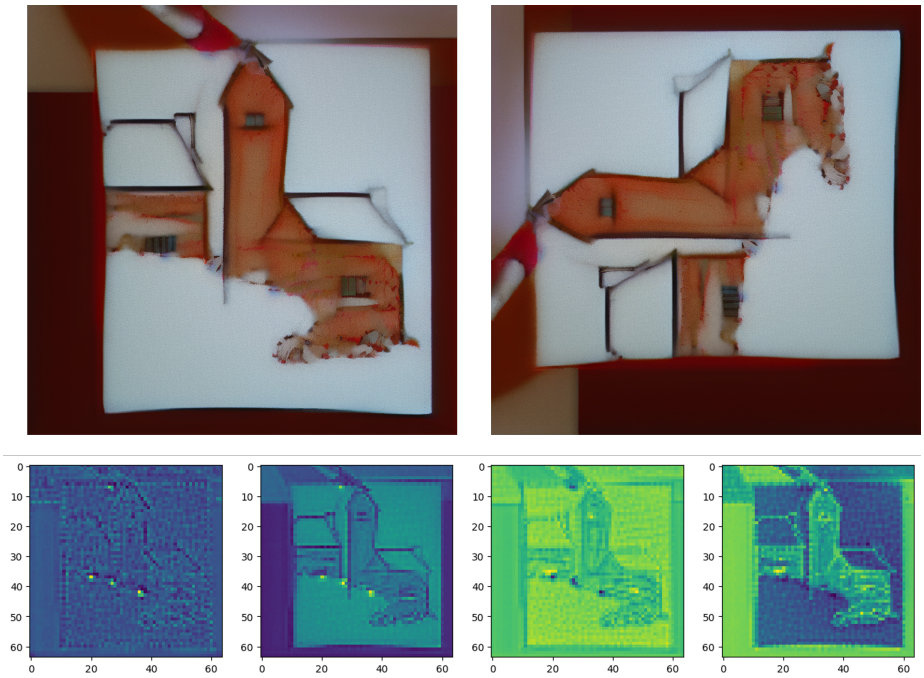


Figure 14: Generation with text prompts "A cartoon drawing of a horse" and "A cartoon drawing of a snowy mountain village" after applying motion illusion filter to latent channel 1. And latent space visualization of each channel

7.3 Problematic Artifacts Assumption

Second, we naively draw an equivalence between latent-based artifacts, noise, and rich texture. The botanical garden generation is a typical example that reflects this problematic assumption: the rich textures are simultaneously removed when artifacts are mitigated by the image processing block.

7.4 Lack of Analysis of Latent-based Artifacts

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, more comprehensive research is required on the cause of this artifact, and probably adding priors from other modalities can help with fixing it. For instance, we experiment with generating the visual anagram from the same text prompt with and without a sounding prior. The generation results are illustrated in Fig 15. With the sound prior, we argue that both views look more reasonable than the generation without it.

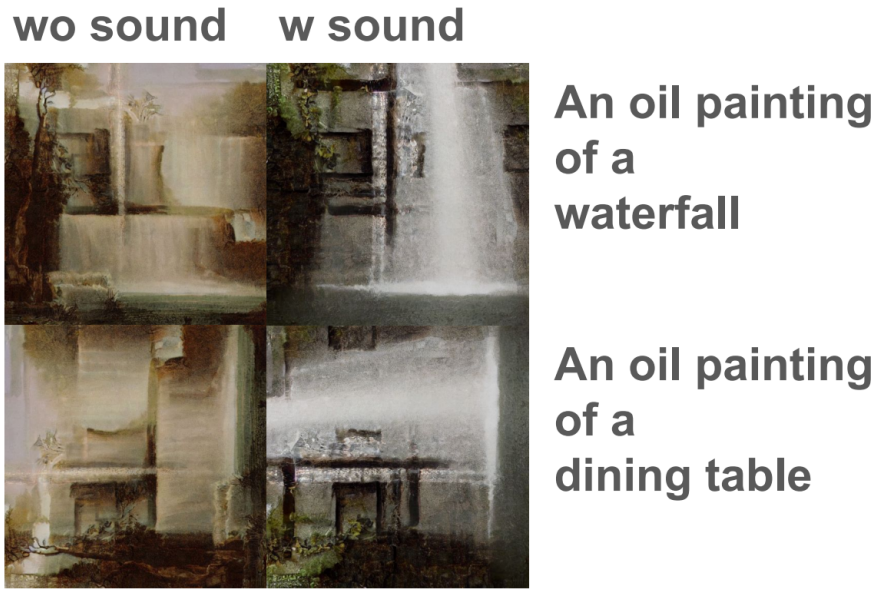


Figure 15: Comparison between raw generations using the same text prompts with and without sound priors.

8 Acknowledgements

Acknowledgements We appreciate the discussion with Professor Andrew Owens, Professor Liyue Shen, and Jason Hu. They gave us valuable suggestions and feedback.

9 Code

https://github.com/scottyehengz/EECS-556-Project

References

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

[2] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017. https://ieeexplore.ieee.org/document/8022901.

[3] PIZER SM. Adaptive histogram equalization and its variations. *Computer Graphics and Image Processing*, 6:184–195, 1987. https://www.sciencedirect.com/science/article/abs/pii/S0734189X8780186X.

[4] Yang Wang and Zhibin Pan. Image contrast enhancement using adjacent-blocks-based modification for local histogram equalization. *Infrared Physics & Technology*, 86:59–65, 2017. https://www.sciencedirect.com/science/article/abs/pii/S135044951730155X.

[5] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. Recent progress on generative adversarial networks (gans): A survey. *IEEE access*, 7:36322–36333, 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=8667290.

[6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. https://arxiv.org/pdf/1606.05908.

[7] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. https://arxiv.org/pdf/2107.00630.

[8] Akira Kubota, Aljoscha Smolic, Marcus Magnor, Masayuki Tanimoto, Tsuhan Chen, and Cha Zhang. Multiview imaging and 3dtv. *IEEE signal processing magazine*, 24(6):10–21, 2007. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=4388078.

[9] William T Freeman, Edward H Adelson, and David J Heeger. Motion without movement. *ACM Siggraph Computer Graphics*, 25(4):27–30, 1991. https://dl.acm.org/doi/pdf/10.1145/127719.122721.

[10] 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. https://arxiv.org/pdf/2103.00020.