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

March 29, 2024

1 Honor Code

Our team will give attribution for any figures used in our documents and will cite all code sources.

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. [1]. But 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 [1]. 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 [2] 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 [3] 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) [4], Variational Autoencoders

(VAEs) [5], and Diffusion Models [6], 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 be obscured or less noticeable in traditional 2D images pose obstacles to comprehend the entire spectrum of features within the image [7]. The paper [8] 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 [8]. 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 [8] 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 [8] 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 be obscured or less noticeable in traditional 2D images.
- Combined diffusion model and traditional image processing tools to generate diverse and high-quality images from limited data.
- We provide quantitative and qualitative results to demonstrate both the efficiency and flexibility of our method.

3 Quantitative Performance Prediction

3.1 Time of generating images

We plan to measure the time of generation for images with shapes 64×64, 256×256, 1024×1024 to demonstrate we speed up the process by replacing

the pixel-based diffusion model with a latent diffusion model(LDM). We predict that our optimized generative framework, incorporating Fourier (or other) Feature Networks and advanced diffusion models, will significantly reduce the image generation time compared to baseline generative models. This prediction is testable by measuring the wall-clock time of generating a predefined set of complex illusions, comparing our method against standard diffusion models without our optimizations.

3.2 CLIP Evaluation

We plan to use the CLIP [9] as part of the metric for quantitatively evaluated our results. CLIP measures how well views align with the text prompt by calculating the similarity between image embedding and textual embedding. The score matrix $S \in \mathbf{R}^{N \times N}$ is defined as:

$$S_{ij} = \phi_{img}(v_i(x))^T \phi_{text}(p)j),$$

where ϕ_{img} and ϕ_{text} are the CLIP visual and textual encoders respectively. Both encoders return a unit-norm vector. x is the generated illusion, and v_i are the views with associated prompt p_i. The dot product is used here, and a higher dot product of embedding implies a higher similarity between the image and the text. We expected that as our method speeds up the process using Latent Diffusion Models(LDMs), it could also match or surpass the CLIP evaluation of Geng et al [8].

3.2.1 Alignment

The first metric we used is the alignment score $\mathcal{A} = \min \text{diag}(S)$ [8]. The alignment score $\mathcal{A}$ intuitively measures the worst alignment of all the views. We forecast an improvement in the alignment score over existing methods, as quantified by the cosine similarity between image embeddings and text embeddings extracted from a pre-trained CLIP model. This metric evaluates how well the generated images align with textual prompts, indicating the effectiveness of our text-to-image synthesis.

3.2.2 Concealment

However, the alignment score $\mathcal{A}$ does not account for the possibility of seeing prompt p_i in view v_j for $i \neq j$. To quantify these occasional failure cases, the concealment score is defined as $\mathcal{C} = \frac{1}{N}\mathbf{trace(softmax}(S/\tau\mathbf{))}$ [8], where τ is the temperature parameters of CLIP. In computing the concealment score $\mathcal{C}$, both directions of the softmax are averaged, and hence this metric measures how well CLIP can classify a view as one of the N text prompts and vice versa.

3.3 Subpredictions

Subpredictions include improvements in aesthetics and diversity scores, drawing from the Aesthetics Predictor and Vendi Score metrics [10], respectively.

These enhancements are anticipated due to our method's advanced handling of high-frequency image features and efficient parameter optimization. This outlines an experimental framework where these predictions can be quantitatively evaluated using a dataset of images generated under controlled conditions. We believe that our approach will not only validate the proposed predictions but also contribute significantly to the fields of visual illusion generation and computational creativity.

4 Plan

Model Setup and Reproduction We first want to reproduce the results shown in the Visual Anagrams model to generate images that have multi-view optical illusions with the published code. [8] We will start with the dataset that the authors compiled by hand and then generalize to the 10 classes from CIFAR-10 which contains a prompt per pair of classes. [11]

We will perform a test run of this model with computing resources provided by this class and record the time of generating a batch of images. This will enable us to evaluate the efficiency of the original framework, which will be used as a benchmark to assess our methods for speeding up the image generation process. Four of us will work on this step together so that we can all comprehensively understand the details of the model. We will presumably finish this step in a week.

Method: 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 [12] as a constraint to eliminate the artifact from the latent domain.(**Figure 1**) 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 (Xingjian and Xiaofeng will work on flip and Wonseok and Yeheng will work on rotation).

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 (add motion illusion as another constraint in the latent diffusion model) with their pixel-based diffusion model in the aspect of how the results correspond to the text prompt and time efficiency. We would expect our model to significantly speed up the image generation process while having a similar performance on the multi-view optical illusion and prompt alignment.

Alternative Method and Post-generation In case of failure of this motion illusion method, we will parallelly experiment with our alternative plan,

which is designing a decoder and invertible transformation for the image before passing the diffusion model and accordingly inversely transforming the output and encoding the result as the finalization.(**Figure 1**) This plan essentially relies on the selection of a decoder and transformation. Yeheng and Xiaofeng will experiment with the decoder/encoder design, while Wonseok and Xingjian will test with various transformations. After having some preliminary results, we can further try different combinations between decoder/encoder designs and transformations and fine-tune the parameters, and if this method works, we will also compare the time efficiency and degree of alignment with the original model.

We will largely focus on the motion illusion method by working together while making efforts on the decoder & transformation at the same time as individuals.

If the steps mentioned above went smoothly, we will study the post-generation process on the output images by implementing methods learned from the class. We aim to exact latent information of images that have optical illusions by traditional image processing methods and apply them to make the artificially generated image more natural.

5 Extension

5.1 Speed up the process

Previous preliminary work performed multi-view denoising using Stable Diffusion [13], a latent diffusion model. However, the latent representation effectively encoders patches of pixels. This leads to artifacts under rotation or flips, where the location of latent changes, but the content and orientation of these blocks do not. Although they fixed this issue by implementing their method on a pixel-based diffusion model, the computational costs are significantly higher than the one of latent diffusion models. In our project, we plan to explore the possibility of replacing the pixel-based diffusion model with a latent one. Working directly on the compression level of an image decreases the computational cost as the dimensions of latent features are much smaller than the one of images and meanwhile provides more faithful and detailed reconstructions. To ameliorate the artifacts under rotation or flips when using latent diffusion models, we proposed two plans.

5.1.1 Option 1: add motion illusion as constraint

Freeman et al. [12] described a method for assigning perceptual motion to objects that remain in a fixed position by applying local filters and continuously varying their phases over time. This technique relies on the fact that local phase changes can be interpreted as global movements. By adding this motion illusion as a constraint [12], i.e., let the flipped image match a flipped motion illusion, we hope the content and orientation of each patch change as the location of

latent changes when performing rotation or flips. For each image, we can obtain its motion illusion with the method Freeman et al. [12] described. Since all the transformations we plan to study are orthogonal transformations, we can hardcore these transformations to obtain the transformed motion illusion and use them as guidance.

5.1.2 Option 2: using a dual encoder and decoder

Another plan we came up with is adding a decoding block prior to the latent diffusion model and subsequently adding an encoding block after the latent diffusion model.
For the decoding block prior to the latent diffusion model, we will first pass the image x to a decoder and then apply the transformation v on the output of the decoder. After having the decoded transformed version of the original image x, we pass it to a latent diffusion model. The correspondent post-processing after the latent diffusion model will first apply the inverse transformation v^{-1} and then pass it to an encoder to get the predicted noise. The decoder and encoder before and after the diffusion model do not have to be a deep neural network. We additionally plan to use the Discrete Fourier transform, Discrete cosine transform, and Discrete wavelet transform as the decoders and their inverse transform as encoders.
With further discussion, we found the dual decoder-encoder architecture seems to be too complicated to analyze and experiment at the first stage. Instead, we would like to remove the additionally decoder-encoder pair and replace it with a single image processing block, which will perform operations on the latent. Here are our current option for this block: Gaussian blurring filter, a combination of filters to force the latent to have smooth instantaneous motion, a block that reconstruct the latent by setting the coefficients of high frequency, obtained through Discrete Fourier Transform, to zero.

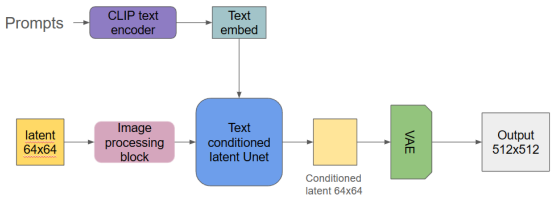


Figure 1: This a simple overview of architecture of a standard latent diffusion model. We neglect the view transformation for simplicity. We hope the imaging processing block can help with mitigating the artifacts in the generation.

6 Progress

6.1 Motion Illusion

Here we applied the method developed by Freeman et al to display patterns that appear to move continuously without changing their positions. [12] 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 filter 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^{theta} with the same method. The visualization of basis is Fig.2:

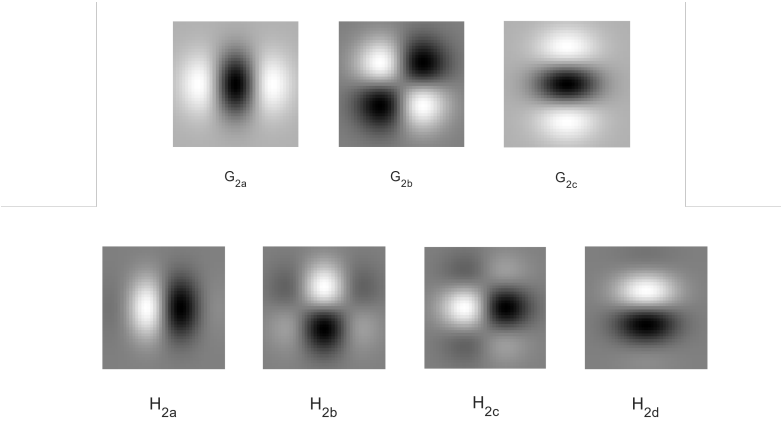


Figure 2: X-Y separable basis filter for G_2 and H_2

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 enable us to calculate image sequence $D(x, y, t)$, where ω is the temporal frequency of the motion. We applied this method to a sample circle image at different angles, to show the performance of this phase-shifted filter on basis element patterns. (Fig.3) We also tested with real images to show the vision

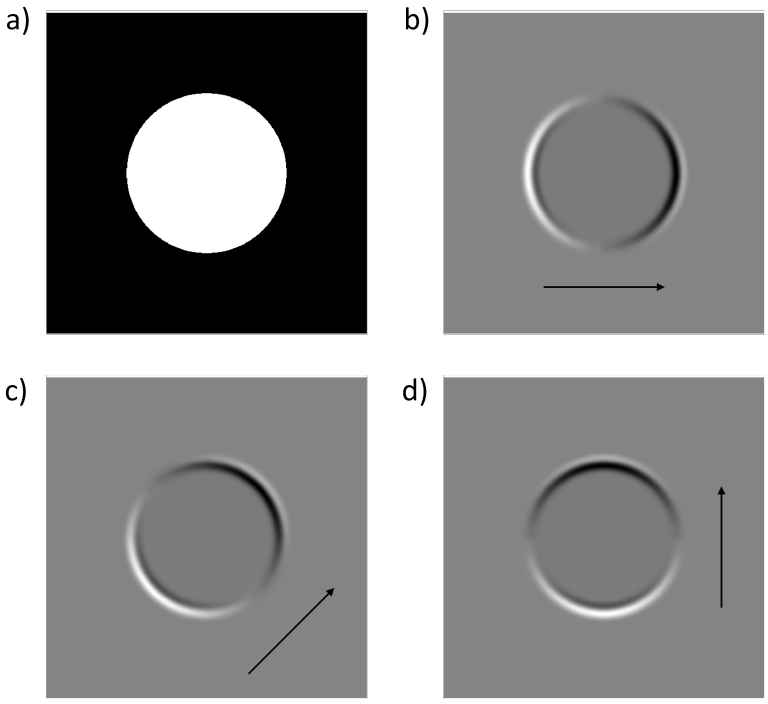


Figure 3: Example of phase-shifted filters applied to a circular disk. a) image of circular disk; b) $\theta = 0°$; c) $\theta = 45°$; b) $\theta = 90°$

illusion of motion as a temporal sequence (Fig.4). To make this applicable to our upgraded multi-view optical illusion generation model, especially the flaws on rational movement. In this scenario, we can determine the orientation of each region, and apply our phase-shifted filters with different directions to the image and assemble the fragments to get the final image. (Fig.5)

We will implement this method to the luminance component of images generated by diffusion model as constrain. By having the rotational phase, we want to test if the machine learning based method can generate images with multi-view illusion whose elements matches better with the rotation operation.

6.2 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.

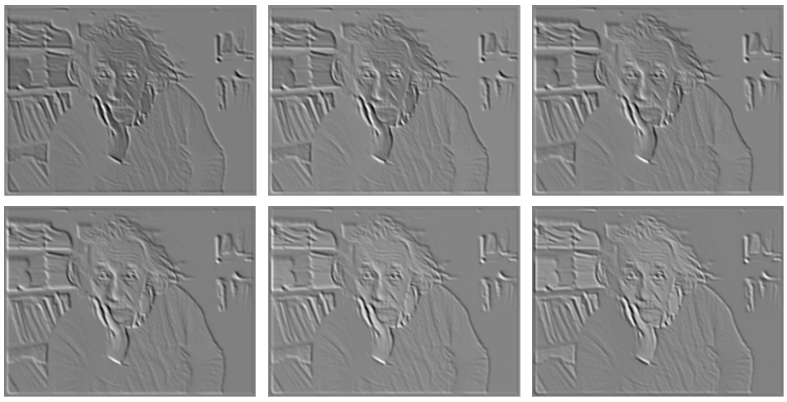


Figure 4: Example of phase-shifted filters applied an image of Einstein. This generates the perception of rightward motion with image remain stationary

The first one is alternatively denoising two views (see Algorithm 1), while the second one is denoising two views in one iteration and then average the predicted noise (see Algorithm 2). For simplicity, we 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.

Algorithm 1 Alternative Denoising

Require: $first\_prompt$, $second\_prompt$, $view$, $inverse\_view$

Prepare text embeddings for both prompts

Prepare scheduler

Prepare latents as random noise ▷ the shape of the latent is typically $1 \times 4 \times 64 \times 64$, corresponds to $B \times C \times H \times W$

for $i, t \in$ enumerate($scheduler.timesteps$) **do**

  if $i\%2 == 0$ **then**

    $latents \leftarrow view(latents, [2, 3])$ ▷ we only want to apply the view transformation on H and W dimension

    $latent\_model\_input \leftarrow$ torch.cat($[latents] * 2$) ▷ concatenate the latents for classifier-free guidance

  else

    $latent\_model\_input \leftarrow$ torch.cat($[latents] * 2$) ▷ concatenate the latents for classifier-free guidance

  if $i\%2 == 0$ **then**

    $noise\_pred \leftarrow$ unet($latent\_model\_input, t, encoder\_hidden\_states = text\_embeddings$).$sample$ ▷ predict noise on one view based on text embeddings of the first prompt

  else

    $noise\_pred \leftarrow$ unet($latent\_model\_input, t, encoder\_hidden\_states = text\_embeddings\_2$).$sample$ ▷ predict noise on another view based on text embeddings of the second prompt

  $noise\_pred \leftarrow$ Apply guidance($noise\_pred$)

  if $i\%2 == 0$ **then** ▷ apply inverse view transform on both the noise and latents to ensure we have the correct updates and denoising the correct latents in the next iteration

    $noise\_pred \leftarrow inverse\_view(noise\_pred, [2, 3])$

    $latents \leftarrow inverse\_view(latents, [2, 3])$

  $latents \leftarrow$ scheduler.step($noise\_pred, t, latents$).$prev\_sample$ ▷ we compute the previous noisy sample $x_t \rightarrow x_{t-1}$

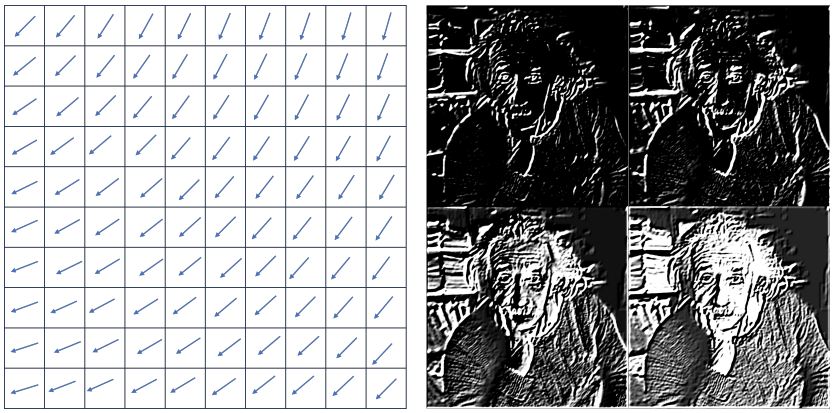


Figure 5: Left. Segments of image regions, and motion orientation of each region from the rotation operation. Right. Rotation motion illusion generated by phase-shifted filter according to the orientations shown on the left. A continuous temporal sequence can be seen in our github link

Algorithm 2 Average Denoising

Require: $first\_prompt$, $second\_prompt$, $view$, $inverse\_view$

Prepare text embeddings for both prompts

Prepare scheduler

Prepare latents as random noise ▷ the shape of the latent is typically $1 \times 4 \times 64 \times 64$, corresponds to $B \times C \times H \times W$

for $i, t \in$ enumerate($scheduler.timesteps$) **do**

$latents\_2 \leftarrow view(latents[2, 3])$ ▷ apply view transform on H and W dimension of latents for the second view

$latent\_model\_input \leftarrow$ torch.cat($[latents] * 2$) ▷ concatenate for classifier-free guidance

$latent\_model\_input\_2 \leftarrow$ torch.cat($[latents\_2] * 2$)

Predict noise for both views

$noise\_pred \leftarrow$ unet($latent\_model\_input, t, encoder\_hidden\_states = text\_embeddings$).$sample$

$noise\_pred\_2 \leftarrow$ unet($latent\_model\_input\_2, t, encoder\_hidden\_states = text\_embeddings\_2$).$sample$

$noise\_pred\_2 \leftarrow inverse\_view(noise\_pred\_2, [2, 3])$ ▷ apply inverse view transformation on H and W dimensions

Perform guidance and combine noise predictions

$noise\_pred \leftarrow$ Apply guidance($noise\_pred$)

$noise\_pred\_2 \leftarrow$ Apply guidance($noise\_pred\_2$)

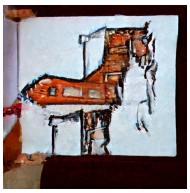
$noise\_pred\_mean \leftarrow 0.5 \times noise\_pred + 0.5 \times noise\_pred\_2$ ▷ average predicted noise from two views

$latents \leftarrow$ scheduler.step($noise\_pred\_mean, t, latents$).$prev\_sample$ ▷ we compute the previous noisy sample $x_t \rightarrow x_{t-1}$

(a) An oil painting of a red panda

(b) An oil painting of kitchenware

Figure 6: An Example of Vertical Flip Visual Anagram



(a) A horse

(b) A snowy mountain village

Figure 7: An Example of Rotation Visual Anagram in Cartoon Drawing Style

We present two figures (Figure 6 and Figure 7) of our preliminary generation results. We reproduce the latent-based artifact mentioned in [8]. When using latent diffusion models we see artifacts that straight lines are forced to be thatched under flip and rotation. Encoding image to latent involves convolution operation and hence manipulating the location of the latent representation does not change the orientation of the pixel blocks after de-convolution operation. We aim to mitigate these artifacts by blending in image processing techniques, and we will discussion the chanllenges and further plan in the later section.

6.3 Evaluation on Preliminary Results

We have embarked on the preliminary evaluation of our model, leveraging widely recognized metrics in the domain of diffusion models. Specifically, we employ the **CLIPScore**, **Fréchet Inception Distance (FID)**, **Precision**, and **Recall** as our primary metrics for assessing the model's performance. These metrics serve as pivotal indicators of the generative model's ability to produce high-quality, diverse, and accurately representative samples of the target data distribution.

CLIPScore is designed for image captioning evaluation. It leverages the capabilities of the CLIP model, which has been pre-trained on a vast number of image-caption pairs, to assess the relevance of captions to images without needing human-written reference captions. This approach reflects a more intuitive, reference-free method of evaluating caption quality, closely mirroring the human process of caption assessment. This is very effective in making the evaluation results in our process.

Fréchet Inception Distance (FID) measures the distance between feature vectors calculated for real and generated images. Lower FID scores indicate a closer similarity to the real image distribution, suggesting higher quality generation. The utility of FID lies in its sensitivity to both the nuances of image composition and the diversity of generated samples, making it an indispensable metric for evaluating image synthesis models.

Precision and Recall, in the context of generative models, offer a nuanced view of the model's performance. Precision assesses the quality of generated images by measuring how many of them are indistinguishable from real images, whereas Recall evaluates the diversity of the generated images by determining how well they cover the real image distribution. A balanced high score in both metrics indicates a model capable of generating diverse, high-quality samples that closely mimic the distribution of the dataset.

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 8: 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"

The generated images for each prompt were then subjected to evaluation using our chosen metrics. The preliminary results are summarized in the table below: we are planning to add the evaluation score of the other metrics such as (FID, precision, and recall) in the final report.

Figure 9: Images of Scenario two

Table 1: Preliminary Evaluation Results of Generated 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

6.4 Challenges and Future Plan

Currently we explore motion illusion and latent diffusion model in a parallel way, and next we will focus on merging and combining our work in image processing technique with the work related to latent diffusion model to mitigate the artifacts. The main challenge we are facing now is how to utilize the motion illusion information or other information obtained be image processing techniques in the latent diffusion.

Regarding to mitigate arifacts, we propose the following future plan. Using the motion illusion information as guidance and as constraint are both reasonable, so we plan to experiment with both of them. In the context of using it as a guidance, at each step of denoising, we plan to perform an additional step to let the latent have a smooth instantaneous motion. We still need more exploration to figure out how to use the motion illusion information as a constraint in the sampling process, but our high-level intuition of this process is to let the samples have a smooth instantaneous motion. Denoising might be another reasonable approach to mitigate artifacts. We would like to first try applying Gaussian Blurring filter to each step. Removing coefficients of high frequency in the latent under Discrete Fourier Transform is another option we considered. We will first see the experiments results and then dig in deep into the one with reasonable output, i.e., a generation with less artifacts.

Regarding to the latent diffusion model, we propose the following plan. First, we will implement more view transformation, clean the code to extract helper functions, and use concatenated input to avoid two denoising pathways. Second, we will explore the options for the text encoder and see if a text embedding in higher dimension would lead to an improvements of generation.

References

[1] 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.

[2] PIZER SM. Adaptive histogram equalization and its variations. *Computer Graphics and Image Processing*, 6:184–195, 1977.

[3] Yang Wang and Zhibin Pan. Image contrast enhancement using adjacent-blocks-based modification for local histogram equalization. *Infrared Physics & Technology*, 86:59–65, 2017.

[4] 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.

[5] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[6] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[7] 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.

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

[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] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[11] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5), 2014.

[12] William T Freeman, Edward H Adelson, and David J Heeger. Motion without movement. *ACM Siggraph Computer Graphics*, 25(4):27–30, 1991.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.