Visual Illusion Generation System Based on Diffusion Models

Zihan Yang
EECS, Peking University

Yuming Fang
EECS, Peking University

Yueran Wang
SMS, Peking University

Abstract

For this Computer Vision final project, we selected the topic of Creating Visual Cognitive Illusions for our experimentation.This paper presents an in-depth study on visual illusion generation using diffusion models. Building upon the reproduction and analysis of the existing Diffusion Illusion model, we identified its implementation flaws and theoretical limitations regarding complex geometric transformations and multi-stable perception. To address these issues, we constructed an Enhanced Illusion Generation Framework. By leveraging the generative prior of Score Distillation Sampling (SDS) and integrating a novel set of differentiable transformation operators into the optimization loop, our framework enables the decoupled optimization of image parameters across frequency domains, color spaces, and geometric perspectives. We successfully extended and implemented 8 distinct categories of visual illusions, including multi-scale hybrids, color-channel separation, cross-modal color-grayscale camouflage, and motion-induced parallax.To enhance reproducibility and interactivity, we developed a dedicated project homepage and an online execution environment. Experimental results demonstrate that our method significantly expands the flexibility and quality of generative illusions, providing new technical avenues for interactive art creation and multi-modal information hiding.And our source code is available this website.

1 Introduction

Visual illusions—images designed to deceive the human visual system—encompass various categories, including ambiguous illusions, geometric-optical illusions, and motion illusions. Historically, they have remained a focal point of research in both psychology and computer graphics. In recent years, Diffusion Models have demonstrated exceptional capability in modeling complex distributions. Their scope of application has expanded far beyond mere image synthesis, extending to general robotic action trajectory generation[4] and Sim-to-Real navigation strategies (e.g. NavDP[2]). However, despite the revolutionary breakthroughs Generative AI has brought to image creation, precise control over the geometric features and semantic attributes required for visual illusions remains a significant challenge.

Building upon the "Diffusion Illusion" framework by Ryan[1], we identified theoretical limitations in existing models, specifically their restriction to a narrow range of illusion types and their reliance solely on text-based inputs. To address this, we focus on ambiguous illusions and extend the framework by incorporating an image-guidance mechanism and a suite of differentiable transformation operators. This approach not only enriches the diversity of generatable illusions but also supports user-supplied reference images. Specifically, we have implemented the following effects:

Contributions. Our main contributions are summarized as follows:

- **Reproduction and Enhancement:** We successfully reproduced the three baseline illusions from the original *Diffusion Illusions* framework and introduced critical improvements to their stability and visual quality, addressing limitations in the original implementation.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- **Automated Evaluation Pipeline:** We constructed a novel closed-loop pipeline that integrates automated generation with Vision-Language Model (VLM) based scoring. This system enables the objective quantification of illusion effectiveness and facilitates the high-throughput screening of optimal samples, replacing subjective manual selection.
- **Physical World Validation:** We bridged the gap between digital synthesis and physical perception by printing our generated samples. Experiments confirm that our illusions—particularly the cylindrical anamorphosis and hybrid images—maintain their perceptual efficacy in real-world environments under varying viewing conditions.
- **Framework Abstraction and Extension:** We formalized the previous specific implementations into a **Unified Multi-View Optimization Framework**. Based on this abstraction, we significantly expanded the solution space by introducing eight novel illusion categories, the implementation details of which are comprehensively described in Section 4.
- Finally, we developed a comprehensive homepage with an interactive interface, allowing users to explore our results and generate their own illusions in real-time.

2 Related Work

2.1 Traditional Methods and Controlled Generation

Early illusion generation relied heavily on frequency analysis or manual geometric design. Oliva[10].'s Hybrid Images exploited multi-scale perception to blend frequency bands, while works on Camouflage and Anamorphosis [5] focused on texture optimization and geometric projection. With the rise of generative AI, methods like ControlNet[14] and Illusion Diffusion[1] enabled the embedding of hidden patterns (e.g. QR codes) via spatial conditioning. However, these inference-based approaches often struggle to generate multi-stable illusions that require consistency across varying viewpoints or frequency domains.

2.2 Optimization-based Paradigms

To overcome single-view limitations, optimization-based frameworks leverage Score Distillation Sampling (SDS) to enforce multi-view consistency. Diffusion Illusions[1] pioneered optimizing a single image against prompts under affine transformations. Visual Anagrams[7] extended this using pixel-space models for rearrangement illusions, while DreamFusion[11] applied SDS to 3D generation. Our work builds on these foundations but introduces specialized decoupling operators for frequency and color spaces, addressing the limitations of current methods in handling multi-scale and cross-modal illusions.

3 Data

As our method is a zero-shot optimization approach, we do not require a traditional training dataset. Instead, we utilize:

- **Pre-trained Model** Stable Diffusion (based on the work of Diffusion Illusion[1]) acts as the source of visual knowledge and DeepFloyd IF.
- **Input Prompts** We carefully designed a series of prompts with striking contrasts for illusion generation, which yielded favorable results. In addition, we incorporated images generated by nanobanana as prompt inputs and embedded them into the illusions.
- **Evaluation** For quantitative assessment, we grasp idea from IllusionVQA[12] and GVIL[13] datasets.

4 Method

4.1 Unified Multi-View Optimization Framework

This study is deeply inspired by "Diffusion Illusions" proposed by Burgert[1] Through an in-depth analysis of the original code, we abstract and formalize the generation process of such visual illusions into a **Unified Multi-View Optimization Framework**.

Algorithm 1 Unified Framework for Diffusion-Based Visual Illusions

Require: Pre-trained Diffusion Model ϵ_ϕ, Set of Transformations $\mathcal{T} = \{T_1, \ldots, T_K\}$, Set of Targets $\mathcal{Y} = \{y_1, \ldots, y_K\}$, Total iterations T

Initialize: Learnable parameters θ (e.g., Fourier features, voxel grid, or texture layers)
Optimizer: Initialize optimizer (e.g., Adam, SGD) for θ
for $t = 1$ to T **do**
  $x \leftarrow \text{Construct}(\theta)$ {Synthesize base object from parameters}
  $\mathcal{L}_{\text{total}} \leftarrow 0$
  for $k = 1$ to K **do**
    $v_k \leftarrow T_k(x)$ {Apply differentiable transformation}
    if y_k is a Text Prompt **then**
      $\mathcal{L}_k \leftarrow \mathcal{L}_{\text{SDS}}(v_k, y_k, \epsilon_\phi)$
    else if y_k is a Target Image **then**
      $\mathcal{L}_k \leftarrow ||v_k - y_k||_2^2$
    end if
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{total}} + \lambda_k \cdot \mathcal{L}_k$
  end for
  $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{total}}$ {Update parameters via backpropagation}
end for
return Optimized object x

We parameterize the object to be optimized as θ (which can be a single image, multi-layer textures, or a 3D voxel grid). We define a set of differentiable transformation functions $\mathcal{T} = \{T_1, \ldots, T_K\}$, where each T_k represents a specific physical observation modality or image processing operation. Our objective is to find a parameter set θ such that all transformed views $v_k = T_k(\theta)$ simultaneously satisfy their respective target constraints y_k.

The general pseudocode for this framework is presented in Algorithm 1.

4.2 Baseline Implementation

We first summarize the three baseline implementations (visualized in Figure 3) by Burgert[1], which constitute special cases of this framework:

- **Flippy Illusion:** θ is a single image. The transformation T is a 180° rotation. The image conforms to two distinct text descriptions when viewed upright and inverted.
- **Rotation Overlays:** θ contains top and bottom image layers. The transformation T is defined as the superposition product of the two layers, where the top layer is rotated by multiples of 90°.
- **Hidden Characters:** θ consists of four independent images. The transformation T is the pixel-wise product of the four images, revealing hidden information upon superposition.

4.3 Our Improvements

Based on the unified framework described above, we expanded the boundaries of the transformation function $T(\cdot)$ and proposed eight novel visual illusion applications.

1. **Image-Driven Hard Constraint Optimization:** Addressing the limitation of the original framework which relied solely on stochastic text prompts (via SDS Loss), we introduced a deterministic image-guidance mechanism. By incorporating a pixel-wise Mean Squared Error (MSE) loss into the optimization loop, we allow the system to use user-provided images as rigid targets. This enables precise control over the generated illusion, such as forcing the superposition of multiple layers to reconstruct a specific logo or reference image rather than a generic semantic concept (Figure 1a).
2. **Multi-Angle Moire Cryptography:** Unlike the rigid 90-degree rotation in the baseline, we introduce arbitrary affine transformations combined with specific target MSE constraints. We optimize a dual-layer texture structure (Base Layer and Decoder Layer) such that when

superimposed at specific angles ($0°, 120°, 240°$), high-frequency texture interference (Moiré effect) accurately decodes distinct hidden visual symbols, achieving rotation-angle-based visual decryption (Figure 1b).

3. **Differentiable Cylindrical Anamorphosis:** We design T as an optical mapping function simulating cylindrical reflection. Using differentiable Grid Sample techniques, we map the planar canvas to the cylindrical mirror coordinate system. During optimization, the original canvas is constrained to exhibit abstract textures formed by stretching, while the reflected image after T_{mirror} transformation is constrained to exhibit concrete target images. This achieves inverse rendering of physical optics (Figure 1c).
4. **Motion Integration Steganography:** Leveraging the persistence of vision effect, we define T as the directional motion blur convolution kernel. The optimization goal is to make the static image present a visible high-frequency texture, while the blurred image after rapid shaking (simulated convolution) reveals a hidden target. This effectively encodes target information into the image's noise patterns, which is decodable only under temporal integration (Figure 1d).
5. **Orthogonal Voxel Projection Synthesis:** We elevate θ to a 64^3 resolution 3D Voxel Grid. The transformation T is defined as volume projection (Mean Projection) along the three orthogonal axes X, Y, and Z. This allows us to decompose the 3D generation problem into three 2D diffusion guidance problems, thereby generating an "impossible" 3D structure that presents completely different objects from different orthogonal perspectives (Figure 1e).
6. **Distance-Dependent Spectral Hybridization:** Inspired by Hybrid Images theory, we explicitly decompose θ into multiple frequency layers. T is a set of Gaussian filters with different σ (low-pass, band-pass, high-pass). We separately optimize low-frequency (far view), mid-frequency (mid view), and high-frequency (near view) components to correspond to different semantic descriptions. This achieves continuous semantic transition based on viewing distance within a single image (Figure 1f).
7. **Cross-Domain Luminance Decoupling:** Utilizing the orthogonality of color perception, we define T as the luminance conversion formula from RGB to grayscale ($Y = 0.299R + 0.587G + 0.114B$). We simultaneously impose uncorrelated semantic constraints on the original color image and the transformed grayscale image. Through optimization, we generate visual adversarial images that present one object in color mode but abruptly mutate into another in desaturated (black and white) mode (Figure 1g).
8. **Intra-Channel Frequency Splitting:** We refine the frequency separation strategy into the RGB channels. The transformation T extracts high and low-frequency components for each of the R, G, and B channels. This allows us to encode non-interfering visual information separately within the macro color block distribution (low frequency) and micro detail textures (high frequency) of the same color image, achieving finer texture nesting than traditional hybrid images (Figure 1h).

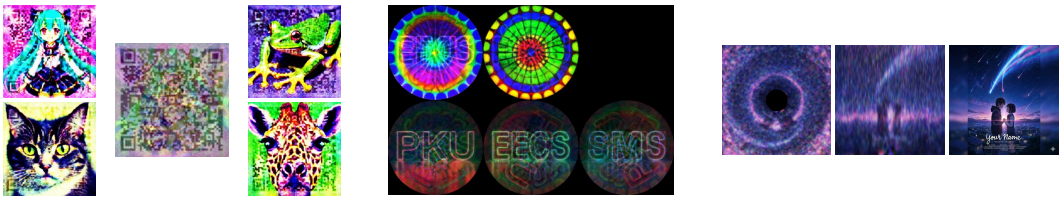
5 Experiments

5.1 Experimental Settings

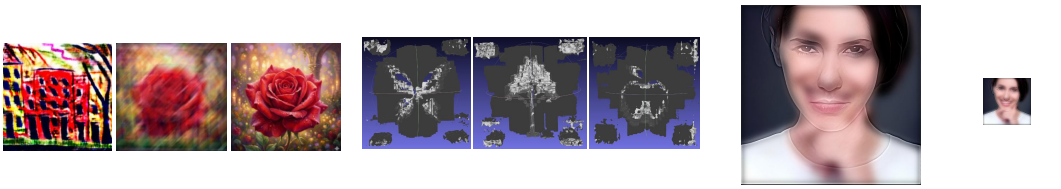
To evaluate the quality of visual illusions generated by Diffusion Models, one viable approach is to employ Vision Language Models (VLMs) to verify whether the generated illusion images convey the intended semantics. Several studies have explored this direction, such as *IllusionVQA* [12] and *GVIL* [13].

In this assignment, we designed an automated pipeline for the generation and evaluation of visual illusions. The framework of this pipeline consists of two main stages: automated generation using Diffusion Models and intelligent evaluation using VLMs. Under this framework, we evaluated and analyzed two types of illusions:

1. **Text Flip Illusion (Ambigrams):** An image that reveals one string of text when viewed upright and a different string when rotated by 180 degrees [7].
2. **Hybrid Illusion:** As described in Section 4, where perception changes based on frequency or color modality.



(a) Hidden QR Code (b) RGB Moire Cryptography (c) Cylindrical Anamorphosis



(d) Motion Steganography (e) Orthogonal 3D Projection (f) Hybrid Expression

(g) Color-Grayscale Switch (h) Color Temperature Shift

Figure 1: **Gallery of Generated Illusions.** We demonstrate eight distinct visual illusion categories enabled by our unified framework. (a) Four images (Miku, Frog, Cat, Giraffe) superimposed to form a functional QR code for our project homepage. (b) Two layers reveal a red "PKU", green "EECS", and blue "SMS" logo when superimposed at $0°$, $120°$, and $240°$ respectively. (c) (Left) The planar "starry sky" canvas. (Center) The reflection visible on the cylindrical mirror. (Right) The target image (movie poster from *Your Name*). (d) (Left) Image of buildings. (Center) The effect after simulated shaking (motion blur). (Right) The target image (a rose). (e) A generated 3D asset appearing as a butterfly, a tree, and an apple from three orthogonal viewing directions. (f) (Left) A non-smiling person. (Right) The downscaled version appears to be smiling. (g) (Left) A colorful landscape with a river and mountains. (Right) The grayscale version reveals a tiger. (h) (Left) The image shows cool color tones. (Right) The downscaled version shifts to warm color tones.

5.2 Pipeline Architecture and Implementation

Our experimental pipeline operates in two sequential stages, each equipped with dedicated scripts and producing structured output files. All scripts, experimental results (including generated images), natural language evaluations from the VLM, and numerical scores (stored as JSON files) have been uploaded to our project's GitHub repository.

Stage 1: Automated Generation. For the Text Flip Illusion, we utilized DeepFloyd IF as the backbone model due to its superior character generation capabilities. For the Hybrid/Blurry Illusions, we employed Stable-Diffusion-v1-4. The generation process yields a structured output directory containing all candidate image samples.

Stage 2: VLM-Based Intelligent Evaluation. Each generated candidate image was evaluated using **Qwen3-VL-4B-Instruct**, a Vision Language Model capable of understanding both visual content and text. To ensure objectivity, we implemented a "blind test" strategy where the VLM performs recognition without prior knowledge of the target words.

5.3 Experimental Results

For the Text Flip Illusion, we conducted experiments on two word pairs: (EECS, SMS) and (ICS, LOVE). We generated 3,333 candidate images for the (EECS, SMS) pair and 1,923 candidates for

the (ICS, LOVE) pair. Additionally, for the Hybrid Illusion, we prepared a pool of 20 pre-defined prompts; for each generation, a prompt was randomly selected from this pool, resulting in a total of 100 sets of generated images. Due to space constraints, we present only the most representative experimental results below.

Table 1 illustrates the score distribution for the (EECS, SMS) pair. As observed, generating high-quality visual illusions of this type presents significant difficulty, with only a very small fraction of samples aligning well with the prompt. Figure 2 (below) displays three sets of generated images which received VLM scores of 0, 5, and 9 respectively, clearly demonstrating the disparity in generation quality.

Table 1: VLM Score Distribution for (EECS, SMS) Illusion Generation

VLM Score	Count	Percentage	VLM Score	Count	Percentage
0.0	1,171	35.1%	5.0	68	2.0%
0.5	36	1.1%	5.5	15	0.5%
1.0	143	4.3%	6.0	106	3.2%
1.5	50	1.5%	6.5	28	0.8%
2.0	77	2.3%	7.0	193	5.8%
2.5	26	0.8%	7.5	666	20.0%
3.0	75	2.3%	8.0	462	13.9%
3.5	15	0.5%	8.5	181	5.4%
4.0	58	1.7%	9.0	4	0.1%
4.5	13	0.4%			

(a) Score 0 Sample (b) Score 5 Sample (c) Score 9 Sample

Figure 2: Comparison of generated samples with VLM scores. (a), (b), and (c) show examples with scores of 0, 5, and 9 respectively. Higher scores indicate clearer text legibility and better illusion consistency.

6 Conclusion

In this project, we established a unified framework for visual illusion generation based on Score Distillation Sampling (SDS), significantly deepening our understanding of controllable generation within diffusion models. By injecting a novel set of differentiable transformation operators into the SDS optimization pipeline, we transcended the limitations of simple 2D rotations. Our work successfully leveraged the nuances of human perception—specifically the interplay between frequency, luminance, and chrominance—to realize **eight** distinct categories of illusions.

While achieving these results, we also identified that Stable Diffusion is lack of Rotation Equivariance. Future work could address this by adopting pixel-space diffusion models (e.g. DeepFloyd IF) or integrating Visual Language Models (VLMs) for automated illusion evaluation. Detailed documentation and a more comprehensive analysis of the limitations are provided in the Appendix.

To facilitate reproducibility and interaction, we developed a project homepage and a Google Colab environment, allowing users to reproduce all generated illusions.

References

[1] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery.

[2] Wenzhe Cai, Jiaqi Peng, Yuqiang Yang, Yujian Zhang, Meng Wei, Hanqing Wang, Yilun Chen, Tai Wang, and Jiangmiao Pang. Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance. *arXiv preprint arXiv:2505.08712*, 2025.

[3] Pascal Chang, Sergio Sancho, Jingwei Tang, Markus Gross, and Vinicius C. Azevedo. Lookingglass: Generative anamorphoses via laplacian pyramid warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[5] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010.

[6] Ana Dodik, Isabella Yu, Kartik Chandra, Jonathan Ragan-Kelley, Joshua Tenenbaum, Vincent Sitzmann, and Justin Solomon. Meschers: Geometry processing of impossible objects. *ACM Transactions on Graphics (TOG)*, 44(4):1–10, 2025.

[7] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024.

[8] Frederick A. A. Kingdom. Shadows and checkerboards: Quantifying the breakdown of brightness constancy. *arXiv preprint arXiv:2005.08772*, 2020.

[9] Ilia Kulikov, Wang Yifan, P. Alviur, et al. Color illusion diffusion: Text-to-image generation of visual illusions. *arXiv preprint arXiv:2407.03152*, 2024.

[10] Aude Oliva, Antonio Torralba, and Philippe G. Schyns. Hybrid images. *ACM Trans. Graph.*, 25(3):527–532, July 2006.

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

[12] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*, 2024.

[13] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? *arXiv preprint arXiv:2311.00047*, 2023.

[14] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023.

A More Results

A.1 Results from the Original paper

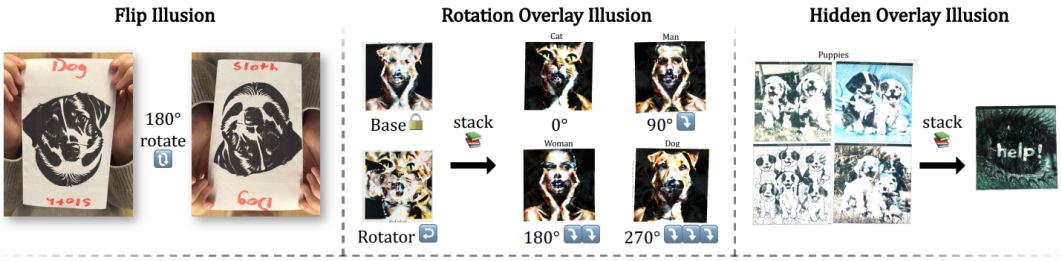
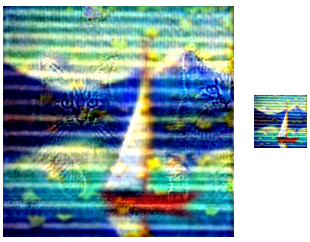
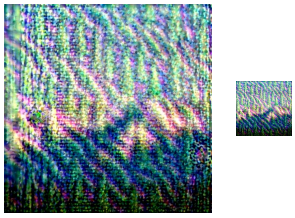


Figure 3: Three optical illusion categories originally proposed by Burgert et al.: the Flip Illusion, the Rotation Overlay Illusion, and the Hidden Overlay Illusion.

A.2 More Results from Our Methods

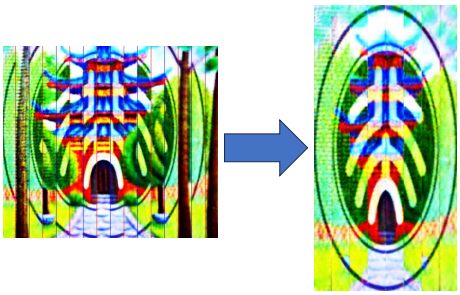


(a) Distance-Dependent Spectral Hybridization

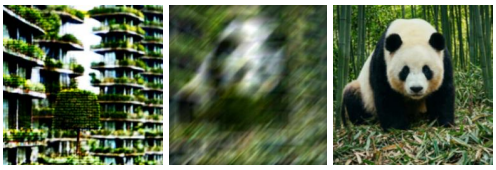


(b) Distance-Dependent Spectral Hybridization

Figure 4: **Distance-Dependent Results.** Examples of hybrid images that change appearance based on viewing distance.



(a) Fold Illusion



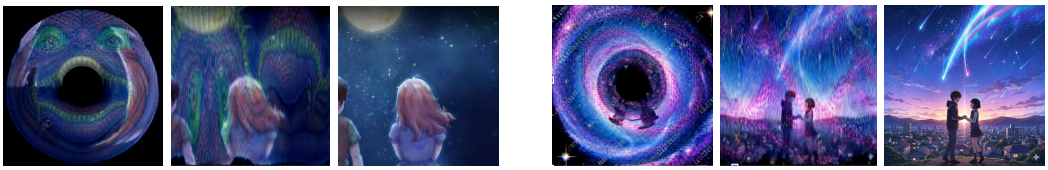
(b) Motion Integration Steganography

Figure 5: **Fold and Motion Illusions.** (a) A preliminary attempt at a **Fold Illusion**. The image depicts a temple; however, when specific sections of the paper are folded inward, the Peking University logo is revealed. (b) An example of motion integration.

A.3 Real World Performances

To validate the physical robustness and practical applicability of our unified framework, we printed the generated illusion patterns onto transparent film and conducted experiments in real-world settings.

As illustrated in the figures below, our method demonstrates promising transferability from the digital domain to the physical world. **Figure** 7 and **Figure** 8 show successful examples where the optical



(a) Differentiable Cylindrical Anamorphosis (b) Differentiable Cylindrical Anamorphosis

Figure 6: **Cylindrical Anamorphosis.** Results showing images that resolve into coherent visuals when reflected on a cylindrical mirror.

illusions remain effective under physical observation conditions. However, we also encountered limitations, as shown in the failure case in **Figure 9**.

This specific failure case was inspired by the concept of "Generative Anamorphoses" proposed in *LookingGlass* [3]. As shown in **Figure 10**, their project page demonstrates compelling visual results using Laplacian Pyramid Warping. However, since the official codebase is unavailable, we attempted to reproduce this effect within our optimization-based framework. While we successfully achieved the target illusion in digital simulations, the physical reproduction failed to reconstruct the target image correctly due to alignment precision and unmodeled physical refraction. Bridging this "Sim-to-Real" gap for complex anamorphic illusions remains a key direction for our future work.

Figure 7: **Physical Validation (Success I).** A successful real-world reproduction printed on transparent film, retaining the illusion effect.

Figure 8: **Physical Validation (Success II).** Another successful example demonstrating the robustness of our method in physical settings.

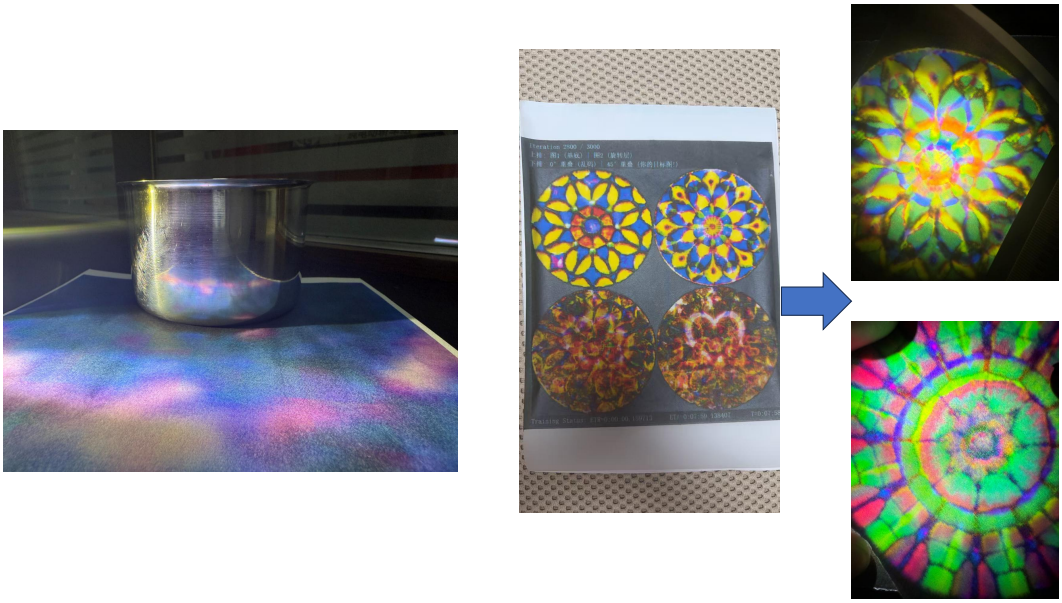


Figure 9: **Physical Validation (Failure Case).** A physical reproduction of the LookingGlass-inspired illusion which failed to reconstruct the target view.

Figure 10: Reference visualization from the *LookingGlass* project page [3], which inspired our attempt at generative anamorphoses.

B Discussion

The analysis above indicates that our pipeline successfully provides an automated method for filtering high-quality visual illusions, effectively validating the scaling law in the context of generative search—where increasing the quantity of candidate generations significantly increases the probability of discovering high-quality outliers.

Furthermore, we offer some insights regarding future directions. A primary challenge in many generative tasks lies in designing the appropriate Loss Function. We hypothesize that VLM automated scoring could potentially serve as a direct Loss Function (i.e., a "Loss Prompt"). However, the current limitation is that such VLM feedback is non-differentiable. If this process could be made differentiable, we believe it would unlock significant applications in optimizing complex semantic constraints.

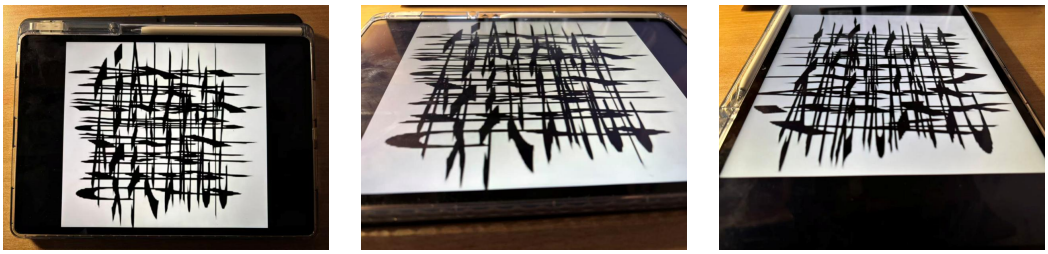
While this study significantly expands the boundaries of diffusion-based visual illusion generation through strategies such as image-driven hard constraints, frequency separation, and cross-domain decoupling, our experiments reveal substantial limitations regarding model architecture, theoretical constraints, and optimization stability.

C Limitations

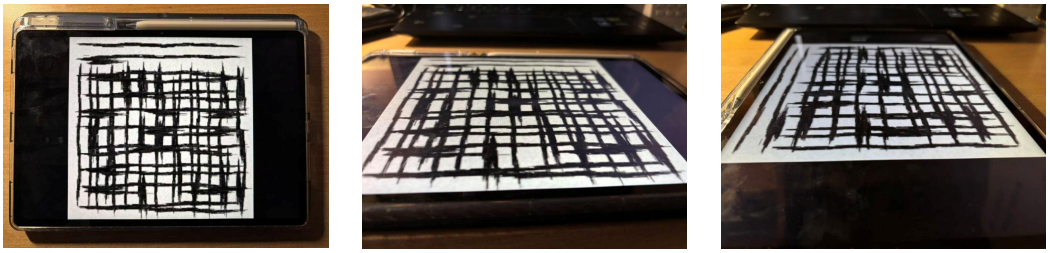
C.1 Limited Typography and Algorithmic Alternatives

Despite the visual fidelity of Stable Diffusion, its capability to generate precise, legible typography remains a significant bottleneck. For geometric illusions requiring exact character alignment and strict perspective consistency—such as the perspective-based text concealment shown in Figure 11—generative approaches often yield illegible glyphs or fail to adhere to the rigid geometric constraints required for the illusion to work.

Consequently, for this specific category, we found that traditional algorithmic generation (deterministic rendering via projection code) significantly outperforms latent diffusion models. As demonstrated in Figure 11(b), our code-based implementation successfully creates a structure that appears as chaotic noise from a neutral viewpoint but clearly resolves into distinct sentences ("I HATE PKU ICS" and "I LOVE PKU CV") when viewed from specific angles. This level of structural precision is currently difficult to achieve using standard Stable Diffusion pipelines due to the lack of explicit character-aware mechanisms in the latent space.



(a) Inspiration: The reference wire-frame text illusion.



(b) Our Algorithmic Generation: Visualizing hidden messages from specific viewpoints.

Figure 11: **Algorithmic Text Illusion Generation.** (a) The reference concept that inspired our experiment. (b) Our reproduction using a deterministic code-based pipeline instead of Stable Diffusion. The structure appears as unintelligible clutter from a normal viewing angle but resolves into the coherent messages "I HATE PKU ICS" and "I LOVE PKU CV" when observed from two specific varying perspectives. This highlights the superiority of algorithmic approaches over diffusion models for precise typographic illusions.

C.2 Artifacts in Latent Space Transformations

Our framework utilizes Stable Diffusion (SD) as the generative backbone, which relies on a Variational Autoencoder (VAE) to compress images into latent space. For tasks involving geometric transformations—such as Orthogonal Voxel Projection or Multi-Angle Moire Cryptography—manipulating the latent grid directly rearranges the spatial configuration of latent patches but fails to rotate the encoded content within the patches themselves. This discrepancy results in "thatching artifacts" (jagged edges at patch boundaries) upon decoding, disrupting the visual continuity of the illusion. Pixel-based architectures, such as DeepFloyd IF, demonstrate superior adaptability to such transformations but entail significantly higher computational costs.

C.3 Gradient Conflicts in Projective Transformations

In exploring Cross-Domain Luminance Decoupling (Color-Grayscale Switch), we encountered challenges more severe than those in traditional geometric illusions. Previous works like Visual Anagrams[7] rely primarily on orthogonal transformations (e.g. rotation), which preserve invertibility and noise statistics. In contrast, the conversion from RGB to grayscale is a non-invertible projective transformation ($\mathbb{R}^3 \to \mathbb{R}^1$). Our experiments indicate that simple channel splicing leads to semantic incoherence, while gradient-based joint optimization suffers from "structural competition." Since the luminance channel carries the majority of structural information, gradients from the color objective and the grayscale objective often conflict violently. This forces the optimization to sacrifice the naturalness of the color image to satisfy the grayscale constraint, frequently resulting in oversaturated artifacts or unnatural textures.

C.4 Trade-off between Pixel Alignment and Semantic Fidelity

Generating visual illusions essentially involves identifying the intersection of multiple semantic concepts on the image manifold. This is particularly challenging when introducing Image-Driven

Hard Constraint Optimization, where we enforce strict pixel-level alignment with reference targets (e.g. specific logos). Such rigid mathematical constraints often force the generative process to deviate from the natural image distribution. Consequently, the optimization tends to compromise texture realism for geometric precision, yielding images that satisfy structural requirements but exhibit perceptual anomalies—such as high-frequency noise or over-sharpening—resembling adversarial examples.

C.5 Over-optimization and Ghosting

Contrary to the expectation that more iterations yield better results, we observed that image quality does not monotonically improve with iteration count in SDS-based optimization. Empirical evidence from our Distance-Dependent Spectral Hybridization experiments suggests that perceptual quality peaks at approximately 1500–2000 iterations. Extending optimization beyond this point (e.g., to the 5000 iterations used in baseline works) leads to "over-optimization," where the model introduces excessive high-frequency details to maximize classifier scores. This results in oversaturation and "ghosting" (where hidden views bleed into the visible view), indicating a lack of a perception-aware stopping mechanism in current algorithms.

C.6 Hyperparameter Sensitivity

By introducing complex transformations in both frequency and spatial domains—such as Intra-Channel Frequency Splitting and Motion Integration Steganography—our method exhibits high sensitivity to hyperparameters. Unlike standard text-to-image generation, illusion synthesis requires finding a Pareto optimal solution among competing loss functions. Critical parameters include spectral cutoff frequencies, loss weights for different views, and the optimization schedule. Experiments show that these parameters often require fine-grained, case-specific tuning to prevent one visual component from dominating the others, hindering fully automated generation.

C.7 Dataset Bias and Generative Boundaries

Our study finds that the generative capacity of diffusion models is significantly constrained by training data priors and encoder architecture, manifesting in three key aspects. First, since SD v1.4 was primarily trained on the LAION-5B dataset, the model exhibits significant Western-centric bias. This leads to severe semantic deviations when processing specific cultural symbols or complex non-Western scenarios, as well as inevitable Western stylistic biases in generated images. Second, the model has internalized strong 3D Consistency Priors of the physical world. When attempting to generate "geometric paradoxes" like Penrose triangles[6], the model tends to use prior knowledge to "correct" spatial contradictions, forcing them into ordinary objects that conform to perspective laws, causing the geometric illusion to fail. Finally, limited by the CLIP text encoder's semantic rather than character-aware mechanism, the model shows instability in handling Typography. When generating text-based illusions such as Ambigrams, the model often treats text as texture rather than symbol sequences, frequently resulting in missing strokes or gibberish, making precise glyph control difficult to achieve.

C.8 Lack of Photometric and Color-based Illusions

While our framework successfully generates structural and multi-view optical illusions, it currently does not account for **photometric or color-based illusions**. Classical examples, such as the Checker Shadow illusion discussed by Kingdom [8], rely on the human visual system's mechanisms of color constancy and brightness assimilation. Our current diffusion guidance focuses on spatial arrangement and semantic consistency, rather than manipulating local contrast to induce color perception biases. Recent work such as *Color Illusion Diffusion* [9] has demonstrated that diffusion models can be adapted to generate such effects via specific attention or brightness guidance. Integrating these perceptual color priors into our pipeline remains a promising direction for future research.

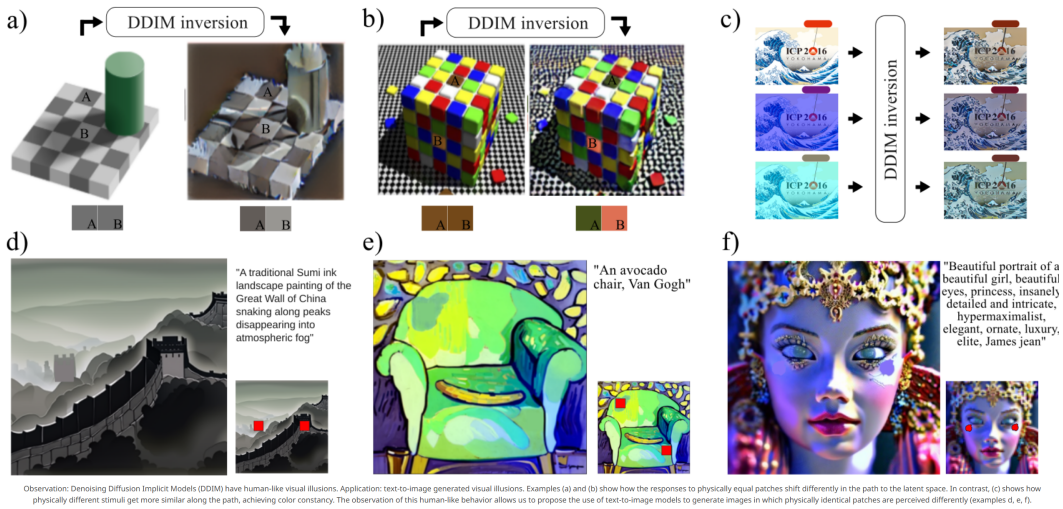


Figure 12: **Photometric Illusions via Diffusion.** Examples from *Color Illusion Diffusion* [9]. Unlike our method which focuses on geometric restructuring, these results demonstrate how diffusion models can be guided to manipulate luminance and color perception (e.g., the cyan and red circles appear different but possess identical pixel values). Incorporating such photometric control is a future direction for our framework.

D Work Allocation

- **Literature Review**: Yueran Wang
- **Codebase Study & Reproduction**: Yuming Fang
- **Idea Investigation & Methodology Research**: Zihan Yang
- **Implementation of Innovations**:
  - First 5 items: Zihan Yang
  - Last 3 items: Yuming Fang
- **Real-world Demonstration**: Yueran Wang
- **Experiment Testing & Evaluation**: Zihan Yang
- **Paper Writing & Formatting**: Yuming Fang, Zihan Yang
- **Homepage & GitHub Setup**: Zihan Yang