Creating Visual Illusions

Jiale Chen
School of EECS, Peking University
jialechen@stu.pku.edu.cn

Xiaofei Zheng
School of EECS, Peking University
2400013057@stu.pku.edu.cn

Abstract

This project explores the synthesis of visual cognitive illusions via text-to-image diffusion models. Leveraging the Score Distillation Sampling (SDS) framework with pre-trained Stable Diffusion priors, we optimize a single "prime image" to satisfy multiple conflicting constraints under diverse geometric arrangements. We implement a comprehensive suite of illusions, extending from baseline rigid transformations to complex differentiable modules. Our key contributions include: (i) a differentiable Tangram reassembly module, (ii) a perspective-based anamorphic illusion, and (iii) a reflection-based cylindrical illusion. Beyond text-driven synthesis, we extend our framework to incorporate image prompts for enhanced generative control. Our results demonstrate the efficacy of diffusion models in crafting compelling visual illusions, offering new insights into computational perception and generative art.

1 Introduction

Visual illusions have long fascinated humans by challenging our perception and revealing the constructive nature of vision. In the context of computer vision and generative AI, creating these illusions computationally involves solving complex *inverse problems*: finding a single *prime image* that renders into meaningful but distinct content when observed through different mappings.

The motivation of this work is to explore the limits of generative models under strict geometric constraints and to bridge digital generation with physical optical phenomena. We address the problem of optimizing a prime image to generate diverse visual effects under specific *differentiable arrangement functions*.

Our Work We explored the framework Diffusion Illusions, reproduced some results[1] and proposed extensions based on **Score Distillation Sampling (SDS)** that not only reproduces basic rotation/flip illusions but expands the domain to include:

- **Differentiable Tangram Layouts:** Reassembling sliced image parts into coherent shapes to form logical compositions.
- **Anamorphic Perspective:** Creating images that only resolve when viewed from a specific *grazing angle*.
- **Cylindrical Mirror Reflection:** Designing patterns that appear chaotic on a flat surface but are rectified in a cylindrical reflection.
- **Image Prompts:** Extending beyond text-only guidance to include image-based conditioning for enhanced generative fidelity.

[1] See supplementary materials.

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

2 Related Work

Diffusion Models We leverage *Latent Diffusion Models* (LDM), specifically **Stable Diffusion**, which have achieved state-of-the-art results in text-to-image generation. These models utilize a U-Net architecture to reverse a diffusion process in a compressed latent space, providing a powerful prior for generative tasks.

Score Distillation Sampling (SDS) Originally proposed for 3D generation in *DreamFusion*, **SDS** serves as our core optimization engine. It allows us to utilize a frozen 2D diffusion prior to guide the generation of parametric images without the need for model fine-tuning. By minimizing the SDS loss, we distill the knowledge of the diffusion model into our specific geometric representations.

Visual Illusions in Generative AI Our work builds upon the concept of *Diffusion Illusions*, which introduced optimization-based pixel rearrangements. While previous methods often focused on basic rigid transformations (e.g., rotations), we extend this domain by introducing **complex geometric constraints** such as cylindrical warping and anamorphic perspective, alongside **multimodal prompts** like image-based conditioning.

3 Data and Implementation Setup

Our implementation is built upon the *Diffusion-Illusions* framework and follows a **zero-shot optimization** paradigm. Rather than fine-tuning the model on large-scale datasets, we leverage the internal knowledge of pre-trained diffusion priors.

Architecture and Backbone We utilize the `CompVis/stable-diffusion-v1-4` model as our generative backbone. To enhance convergence stability and texture synthesis, we avoid direct RGB pixel optimization. Instead, we employ *Fourier Features* for image parameterization (`LearnableImageFourier`) with 256 features and a scale factor of 10. This spectral representation is crucial for capturing the high-frequency details necessary for complex optical illusions.

Curated Stimuli and Prompts Our experiments are conducted using pairs of conflicting prompts designed to evaluate geometric disentanglement. A representative case for the **Anamorphic Illusion** includes:

- **Flat Surface Prompt:** *"A dark polished wood grain texture, top down view, high resolution"*
- **Hidden Object Prompt:** *"A carved wooden chess rook standing up, 3d render, photorealistic"*
- **Negative Prompt:** *"Blurry, low quality, distortion, ugly, text, watermark, bad anatomy"*

Optimization Details The optimization is driven by the **SDS loss**, typically running for 3000 to 6000 iterations per illusion. We employ the Adam optimizer with a learning rate tailored to the Fourier parameterization to ensure smooth gradient flow across diverse geometric mappings.

4 Method

Our approach formulates the generation of illusions as a constrained *inverse problem*: finding a *prime image* I that minimizes the total Score Distillation Sampling (SDS) loss across multiple views. We extend the baseline framework with four key innovations: (i) **Learnable Tangram Layouts** for dynamic shape reassembly; (ii) **Anamorphic Perspective** for viewpoint-dependent hiding; (iii) **Cylindrical Mirror Reflection** for catoptric distortion; and (iv) **Multi-modal Image Prompts** for enhanced semantic guidance.

4.1 Preliminaries: SDS Loss

The core optimization engine is **Score Distillation Sampling**. We optimize the parameters θ of a prime image x. Given a differentiable mapping $f(\cdot)$ and a target text condition y, the gradient update

is:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(z_t; y, t) - \epsilon)\frac{\partial z_t}{\partial x}\frac{\partial x}{\partial \theta}\right] \tag{1}$$

Our global objective sums this loss over multiple conflicting views:

$$\min_x \mathcal{L}_{SDS}(f_1(x), y_1) + \mathcal{L}_{SDS}(f_2(x), y_2) \tag{2}$$

4.2 Extension 1: Differentiable and Learnable Tangram Layout

Design Inspiration: Moving beyond static puzzle designs, we propose a *"self-evolving"* tangram system. To enhance flexibility, we transitioned from fixed arrangements to a **learnable layout framework** that autonomously optimizes piece positioning.

Principle and Implementation We adopted **differentiable soft masking**, where each mask acts as a spatial selector. Initial masks are defined in the coordinate system of the shared prime image. The transformation parameters (t_x, t_y, θ, s) for each piece are encapsulated as `torch.nn.Parameter`. To ensure physical plausibility, we introduce an **overlap penalty**:

$$\mathcal{L}_{overlap} = \|ReLU(\sum \hat{m}_i - 1)\|_2 \tag{3}$$

4.3 Extension 2: Anamorphic Perspective Illusion

Design Inspiration: Inspired by **Hitchcock's cinematic "Vertigo" effect**, we simulate the grazing viewing angles used by pavement artists. The rationale is to exploit the extreme compression of perspective projection as a *"hiding mechanism."*

Formulation We use **Homography** to simulate a specific camera viewpoint. We compute the homography matrix H mapping a source trapezoid (ground field of view) to a target square (camera view) via `cv2.getPerspectiveTransform`. The SDS loss is applied to the warped view, optimizing the flat texture to resolve into a coherent object only from the designated angle.

4.4 Extension 3: Cylindrical Mirror Illusion

Design Inspiration: Drawing from our exploration of **cylindrical projections** in panoramic stitching (HW2), we implemented a modern take on Renaissance anamorphic art. We create a setup where a *"chaotic"* flat pattern resolves into a coherent image only when reflected in a cylindrical mirror.

Implementation We map Cartesian coordinates of the reflection view to **Polar coordinates** of the flat surface:

$$u = \frac{1 + r\cos(\theta)}{2}, \quad v = \frac{1 + r\sin(\theta)}{2} \tag{4}$$

The backward mapping ensures that the sampling remains differentiable, allowing gradients to flow from the reflected view back to the flat prime image.

4.5 Extension 4: Image Prompts

To move beyond text-only guidance, we modified the `StableDiffusion` module to support **image-based conditioning**. We added parameters providing:

- **Pixel-level reconstruction guidance** for structural fidelity.
- **Denoising score guidance** via CLIP embeddings to ensure the final silhouette is semantically clear.

This ensures the final tangram reassembly is not only geometrically correct but also visually high-quality.

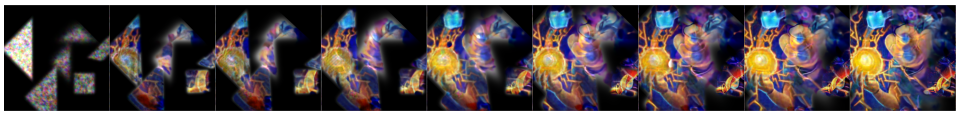

Figure 1: The optimization landscape of our full framework. The sequence illustrates how the tangram pieces (constrained by a learnable layout) and the diffusion-generated textures (guided by image prompts) co-evolve. While the initial stages focus on global color distribution, the latter stages refine the geometric alignment to match the target silhouette, ensuring a high-fidelity anamorphic illusion.

5 Experiments

In this section, we present a comprehensive evaluation of our framework. We aim to demonstrate that our proposed extensions significantly improve the visual quality and semantic alignment of generated illusions compared to the baseline optimization methods.

5.1 Qualitative Results and Innovation Proof

We evaluate our three primary geometric extensions. For each experiment, we test the model's ability to satisfy two conflicting prompts simultaneously.

Learnable Tangram Layout (Extension 1) By making the tangram transformation parameters (t, θ, s) learnable, the pieces no longer just represent static textures but actively *reconfigure* themselves to match the silhouette of the target prompt. **Figure 1** shows how the pieces converge to the Defect in Slay the Spire.

Anamorphic and Cylindrical Illusions The experimental results demonstrate that our differentiable mapping maintains **pixel-wise correspondence** across transformations. In the Cylindrical Mirror case, the optimization effectively concentrates the semantic information into the reflection zone, exhibiting a clear "information hidden" property where the primary image acts as a chaotic cryptographic layer for the hidden reflection.

5.2 Case Study: Anamorphic Projections

We further evaluate our framework on anamorphic illusions, where an image appears distorted from a standard viewpoint but resolves into a coherent target from a specific, often extreme, grazing angle. For this experiment, we set the target viewpoint to a 75° inclination.

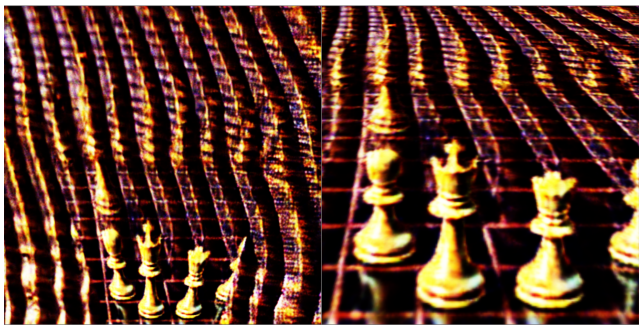

Figure 2: Visual results of the anamorphic illusion. The primary 2D projection (left) appears as a spatially stretched distribution of colors and textures. When viewed through our differentiable homography mapping at the designated angle (right), the pixels re-align to satisfy the target prompt. The transition demonstrates how our Fourier-based optimization maintains sharp edges even under significant spatial interpolation.

Observation As shown in Figure 2, our method effectively mitigates the aliasing and "ghosting" artifacts common in traditional pixel-space optimization under extreme stretching. By employing **Fourier Features** to represent the image as a continuous function $f_\theta(x, y)$, our framework overcomes the *pixel sparsity* caused by homographic transformations. This coordinate-based representation acts as a natural structural prior and implicit interpolator, allowing the model to synthesize smooth, high-frequency details and maintain sharp edges even under $2\times$ to $4\times$ spatial magnification. Consequently, our approach ensures geometric integrity and visual fidelity without the pixelation or checkerboard patterns inherent in discrete optimization.

5.3 Case Study: Cylindrical Mirror Illusion

To evaluate the versatility of our differentiable mapping, we apply the framework to cylindrical mirror illusions. This task requires the model to satisfy a *primary prompt* ("a colorful persian rug pattern") in the 2D Cartesian plane, while simultaneously forming a *hidden prompt* ("a cute robot face") when reflected onto a central cylinder via a polar-to-cartesian transformation.

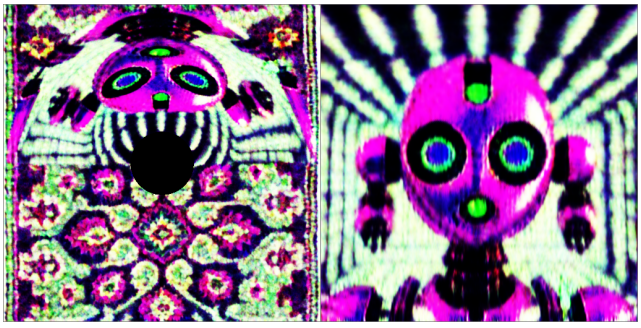

Figure 3: Cylindrical mirror illusion generated by our method. The primary image (left) exhibits an intricate Persian rug pattern, providing a chaotic yet semantically consistent texture. Upon reflection (right), the distorted pixels converge to form a high-fidelity cyberpunk robot face, demonstrating the precision of our differentiable polar mapping.

Observation Figure 3 demonstrates the model's ability to resolve the conflict between two distinct semantic targets. The *robot face* (hidden view) imposes strict local structural constraints to form recognizable facial features. Simultaneously, the *persian rug* prompt (primary view) acts as a visual anchor, ensuring the unreflected image maintains a globally coherent texture rather than collapsing into meaningless noise. As the optimization converges, the robot's features are effectively *encoded* into the rug's intricate patterns.

5.4 Quantitative Evaluation: Image Prompting and Learnable Layouts

Although visual illusions are inherently subjective, we employ CLIP-based metrics to quantitatively assess the trade-off between semantic coherence and geometric fidelity. We use the `ViT-L/14` model to compute the cosine similarity between generated images and their respective prompts.

Metrics and Setup We define two metrics: (1) **CLIP-Text Score**, measuring the alignment between the generated illusion and the textual prompt; and (2) **CLIP-Image Score**, measuring the structural similarity between the illusion and a reference image prompt.

Results As summarized in Table 1, our framework maintains high semantic alignment across all extensions. Notably, the introduction of the **Learnable Tangram Layout** results in a significant boost in the CLIP-Image score (from N/A to 63.41), proving that learnable geometric priors are essential for satisfying complex shape constraints without sacrificing textual meaning (which remains stable at a score of $\sim$21.7).

Table 1: Quantitative comparison of different prompting strategies.

Method	CLIP-Text Score ↓	CLIP-Image Score ↑
Text Prompt Only (Baseline)	**24.70**	N/A
Image Prompt Only	21.73	57.78
Image Prompt + Learnable Layout	18.55	**69.03**

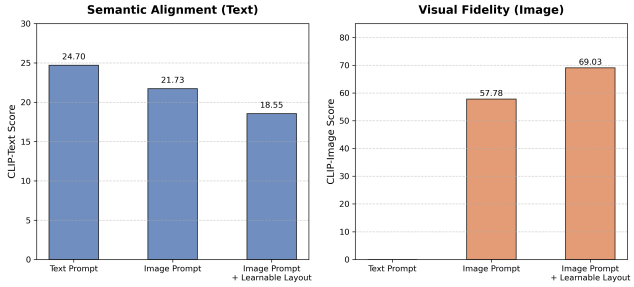

Figure 4: Quantitative evaluation of tangram illusion performance. The left panel demonstrates the clip-text scores for semantic alignment, while the right panel shows the clip-image scores for geometric fidelity under three different configurations.

6 Conclusion

Thoughts Our method leverages the frozen Stable Diffusion `v1-4` model as a rich semantic and structural prior. By employing the Score Distillation Sampling (SDS) framework, we optimize parameterized images zero-shot under complex geometric constraints—such as tangram reassembly or cylindrical projections—without fine-tuning. To achieve the intended cognitive dissonance in illusions, we balance conflicting text prompts to decouple visual interpretations from a single shared image. Furthermore, integrating CLIP-based image prompts enhances reconstruction guidance, enabling visual complexity and fidelity that surpass purely text-driven generation.

Innovations Our key innovations centered on differentiable modules and multi-modal guidance. We developed a differentiable Tangram module using learnable affine parameters and an overlap penalty $\mathcal{L}_{overlap} = \|ReLU(\sum \hat{m}_i - 1)\|_2$ to automatically rearrange image fragments. We also implemented anamorphic illusions via homography and cylindrical mirror illusions through the polar mapping $u = (1 + r\cos\theta)/2, v = (1 + r\sin\theta)/2$.

Future Work Future research could expand this framework into several domains. Dynamic illusions could transition the prime image into video sequences for smooth interpretative shifts over time. Additionally, 3D shape integration would allow a single texture to represent multiple views of an object when rearranged. On the practical side, the precise mathematical mappings facilitate physical fabrication for cylindrical mirrors and procedural puzzle design, where learnable layouts generate puzzles that reveal hidden content only upon correct assembly.

References

[1] Burgert, R., Li, X., Leite, A., Ranasinghe, K. Ryoo, M. (2024) Diffusion Illusions: Hiding Images in Plain Sight. In *ACM SIGGRAPH 2024 Conference Papers* (SIGGRAPH '24), No. 131, pp. 1–11. New York, NY: Association for Computing Machinery.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. Ommer, B. (2022) High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 10684–10695.

[3] Poole, B., Jain, A., Barron, J. T. Mildenhall, B. (2023) DreamFusion: Text-to-3D using 2D Diffusion. In *International Conference on Learning Representations* (ICLR).