

Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models

Daniel Geng, Inbum Park, Andrew Owens
University of Michigan

CVPR 2024

1. Introduction

1.1 What is a Multi-View Optical Illusion?

An image that reveals different content when viewed under different transformations.



an oil painting of people
around a campfire



a lithograph of a table



an oil painting
of an old man

1. Introduction

1.2 Why is Generating Optical Illusions Hard?

1) Multiple Semantic Interpretations

- A single image must represent different concepts in different views.

2) Conflicting Visual Structures

- Different prompts may require incompatible image features.

3) No Explicit Human Perception Model

- Traditional approaches often rely on handcrafted perceptual assumptions.
- Can a pretrained diffusion model discover these illusions automatically?

1. Introduction

1.3 Proposed Solution and Contributions

1) Main Idea

- Instead of generating each view independently, Visual Anagrams performs denoising jointly across multiple transformed views.

2) Core Components

- Multi-View Denosing
- Orthogonal Transformations
- Training-Free Generation

3) Contributions

- A novel multi-view denoising framework for visual anagram generation.
- A theoretical analysis of valid view transformations.
- High-quality multi-view optical illusions generated from pretrained diffusion models.

2. Related Work

2.1 Computational Optical Illusions

Category	Idea	Limitation
Hybrid Images	Frequency blending	Distance-dependent only
Motion Illusions	Phase manipulation	Specific illusion type
Camouflage	Hide objects	Limited flexibility
Bayesian Vision Models	Explicit perception modeling	Requires human vision model

Common Limitation: Most existing approaches rely on handcrafted perceptual assumptions and specialized illusion designs

2. Related Work

2.2 Diffusion models for Illusion Generation

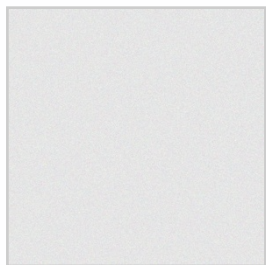
Recent advances in diffusion models provide a new way to generate optical illusions.

Method	Pros	Cons
Diffusion Illusions (Burgert et al., 2023)	- Supports multiple prompts	- Slow SDS optimization - Low output quality
Illusion Diffusion (Tancik, 2023)	- Simple latent-diffusion approach	- Latent artifacts
Visual Anagrams (Ours)	- Fast - Theoretical foundation - Multiple views	- Limited to orthogonal transformations

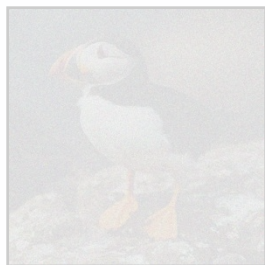
3. Method

3.1 Text-Conditioned Diffusion Models

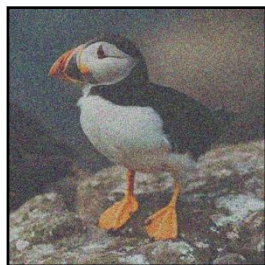
$$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$$



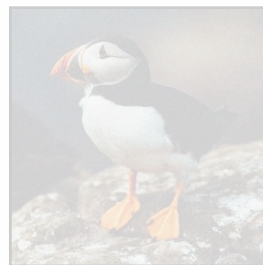
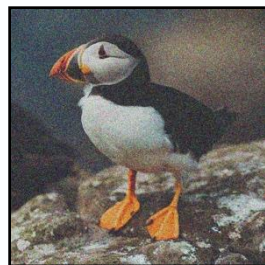
"a photo of
a puffin"



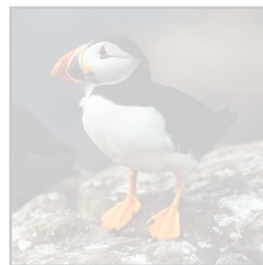
$\mathbf{x}_t$



$\mathbf{x}_{t-1}$



$\mathbf{x}_0$



- x_t : noisy image
- y : text prompt
- t : timestep

Predicted noise $\hat{\epsilon} = \epsilon_{\theta}(x_t, y, t)$

3. Method

3.1 Text-Conditioned Diffusion Models

Classifier-Free Guidance (CFG) $\epsilon_t^{\text{CFG}} = \epsilon_{\theta}(\mathbf{x}_t, t, \emptyset) + \gamma(\epsilon_{\theta}(\mathbf{x}_t, t, y) - \epsilon_{\theta}(\mathbf{x}_t, t, \emptyset)).$

- Unconditional Prediction $\epsilon_{\theta}(\mathbf{x}_t, t, \emptyset)$
- Conditional Prediction $\epsilon_{\theta}(\mathbf{x}_t, t, y)$
- Interpretation $\epsilon_{\theta}(\mathbf{x}_t, t, y) - \epsilon_{\theta}(\mathbf{x}_t, t, \emptyset)$
- Guidance scale γ

3. Method

3.1 Text-Conditioned Diffusion Models

1) Standard diffusion

- One image $\rightarrow$ One prompt $\rightarrow$ One noise estimate $\hat{\epsilon} = \epsilon_{\theta}(x_t, y, t)$ or $\hat{\epsilon}^{\text{CFG}}$

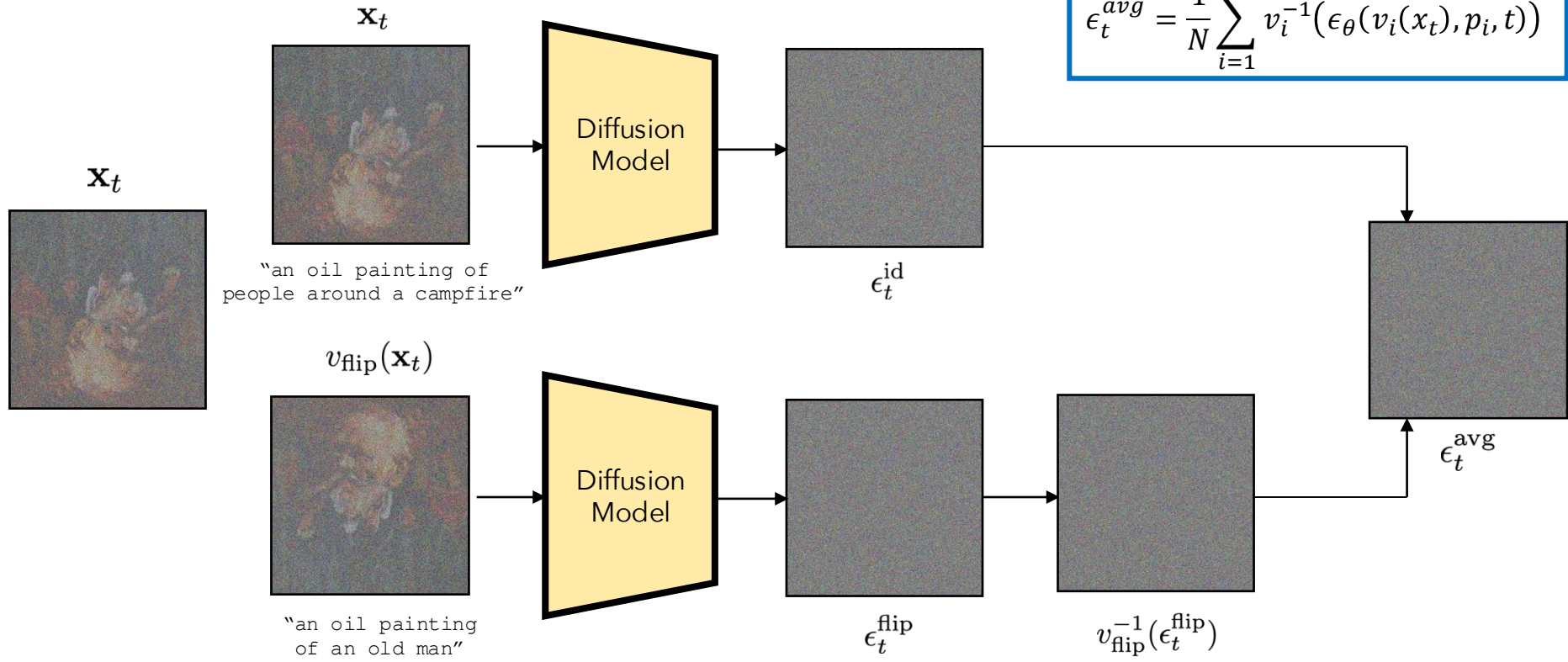
2) Visual Anagrams

- One image $\rightarrow$ Multiple transformed views v_i $\rightarrow$ Multiple prompts p_i $\rightarrow$ Multiple noise estimate

$\rightarrow$ How to combine multiple CFG predictions into a single denoising step?

3. Method

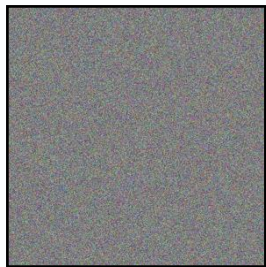
3.2 Parallel Denoising Algorithm



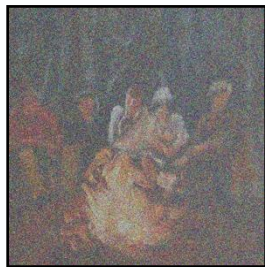
3. Method

3.2 Parallel Denoising Algorithm

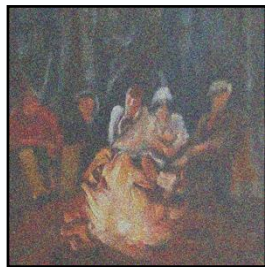
$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$



$\mathbf{x}_t$



$\mathbf{x}_{t-1}$



$\mathbf{x}_0$



$$\epsilon_t^{avg} = \frac{1}{N} \sum_{i=1}^N v_i^{-1}(\epsilon_{\theta}(v_i(x_t), p_i, t))$$

3. Method

3.3 Conditions on Views

A valid view transformation (v_i) must satisfy:

- 1) Invertibility
- 2) Linearity
- 3) Statistical Consistency

3. Method

3.3 Conditions on Views

1) Invertibility

- The transformation must have an inverse.
- Required to align noise estimates back to the original view.

$$\epsilon_t^{avg} = \frac{1}{N} \sum_{i=1}^N \boxed{v_i^{-1}}(\hat{\epsilon}_i)$$

3. Method

3.3 Conditions on Views

2) Linearity

- The noisy data $\mathbf{x}_t = w_t^{\text{signal}} \underbrace{\mathbf{x}_0}_{\text{signal}} + w_t^{\text{noise}} \underbrace{\epsilon}_{\text{noise}}$

$v_i(\mathbf{x}_t) = \mathbf{A}_i \mathbf{x}_t$, for some matrix $\mathbf{A}_i$,
and some flattened noisy image $\mathbf{x}_t$.

$$v\left(\text{img}_1\right) = \boxed{\mathbf{A}} \left(w_t^{\text{signal}} \text{img}_2 + w_t^{\text{noise}} \text{img}_3 \right)$$

$\sim \mathcal{N}(0, \mathbf{I})$

1. Noisy image is weighted sum of *signal* and *noise*

2. Noise must be i.i.d. standard Gaussian

3. Method

3.3 Conditions on Views

2) Linearity

$$\begin{aligned} v_i(\mathbf{x}_t) &= \mathbf{A}_i(w_t^{\text{signal}} \mathbf{x}_0 + w_t^{\text{noise}} \epsilon) \\ &= w_t^{\text{signal}} \underbrace{\mathbf{A}_i \mathbf{x}_0}_{\text{new signal}} + w_t^{\text{noise}} \underbrace{\mathbf{A}_i \epsilon}_{\text{new noise}}. \end{aligned}$$

$$\begin{aligned} \mathbf{x}_t &= w_t^{\text{signal}} \underbrace{\mathbf{x}_0}_{\text{signal}} + w_t^{\text{noise}} \underbrace{\epsilon}_{\text{noise}} \\ v_i(\mathbf{x}_t) &= \mathbf{A}_i \mathbf{x}_t, \text{ for some matrix } \mathbf{A}_i, \\ &\quad \text{and some flattened noisy image } \mathbf{x}_t. \end{aligned}$$

$$v\left(\text{Noisy Image}\right) = w_t^{\text{signal}} \overbrace{\mathbf{A} \text{ (Signal Image)}}^{\text{Transformed Signal}} + w_t^{\text{noise}} \overbrace{\mathbf{A} \text{ (Noise Image)}}^{\text{Transformed Noise}}$$

$\sim \mathcal{N}(0, \mathbf{I})$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

3. Method

3.3 Conditions on Views

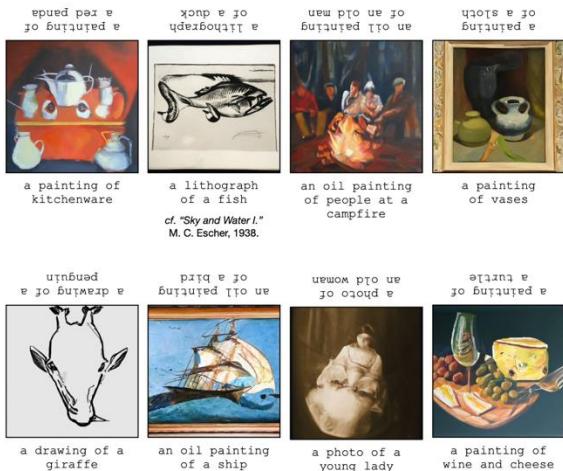
3) Statistical Consistency

- Diffusion models are trained with the assumption that the noise $\epsilon \sim \mathcal{N}(0, I)$
 - Therefore we must ensure that the transformed noise also follows these statistics.
That is, we need $\mathbf{A}\epsilon \sim \mathcal{N}(0, I)$
 - For linear transformations, this is equivalent to the condition that square matrix $\mathbf{A}$ is orthogonal $A^T A = I$
- **A valid view transformation must be orthogonal**

3. Method

3.4 Views Considered

Flips



Three Views



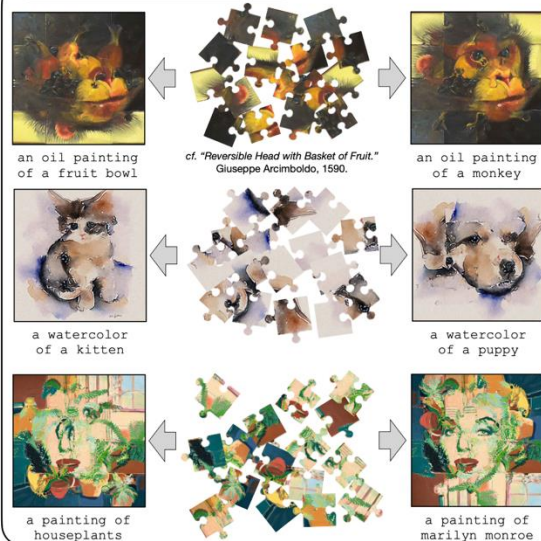
Four Views



Negatives



"Polymorphic" Jigsaw Puzzles



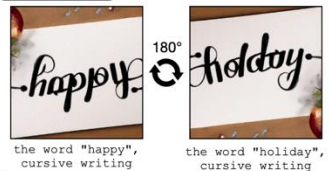
Inner Rotations



Skews



Ambigrams



3. Method

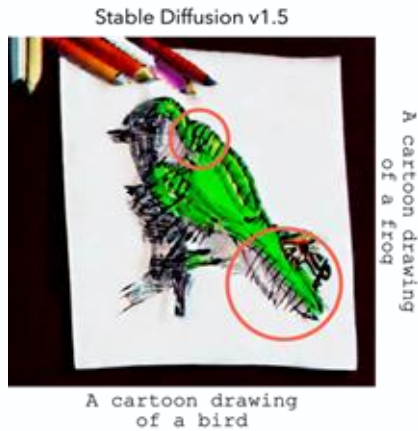
3.5 Design Decisions

1) Pixel Diffusion Model

- Latent models (Stable Diffusion) cause thatched artifacts on rotated blocks. Pixel models (DeepFloyd IF) avoid this.

2) Combining Noise Estimates

- Averaging noise estimates prevents "thrashing" and works better for >2 views compared to alternating timesteps.



4. Experiment

4.1 Quantitative Results (Setup)

1) Baselines

- Burgert et al. (2023) uses Score Distillation Sampling (SDS)
- Tancik (2023) uses latent diffusion

2) Datasets

- CIFAR prompt pairs (45 pairs)
- Custom Prompt pairs (50 manually curated prompt pairs)

3) Evaluation Metric

- CLIP similarity

4. Experiment

4.1 Quantitative Results

Evaluated using CLIP models to measure image-text similarity across views.

1) Alignment Score ($\mathcal{A}$)

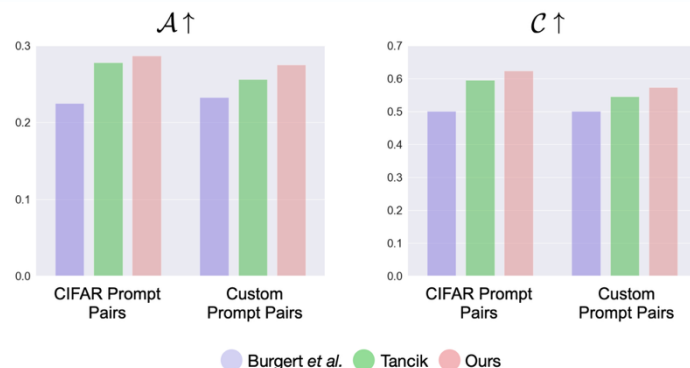
- Measures the worst alignment of all views (min diagonal of score matrix).

2) Concealment Score ($\mathcal{C}$)

- Measures how well CLIP can classify a view as its intended prompt vs. the other views' prompts.

Table 1. **Quantitative Results.** We report the alignment score, $\mathcal{A}$, and the concealment score, $\mathcal{C}$, as well as quantiles of these scores. For a discussion, please see Sec. 4.1.

Prompt Pair	Method	$\mathcal{A} \uparrow$	$\mathcal{A}_{0.9} \uparrow$	$\mathcal{A}_{0.95} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{C}_{0.9} \uparrow$	$\mathcal{C}_{0.95} \uparrow$
CIFAR	Burgert <i>et al.</i> [2]	0.225	0.253	0.260	0.501	0.526	0.537
	Tancik [42]	0.278	0.310	0.316	0.595	0.692	0.712
	Ours	0.287	0.321	0.327	0.624	0.717	0.739
Ours	Burgert <i>et al.</i> [2]	0.233	0.270	0.283	0.501	0.526	0.538
	Tancik [42]	0.256	0.294	0.309	0.545	0.621	0.655
	Ours	0.275	0.315	0.326	0.574	0.668	0.694



4. Experiment

4.1 Quantitative Results

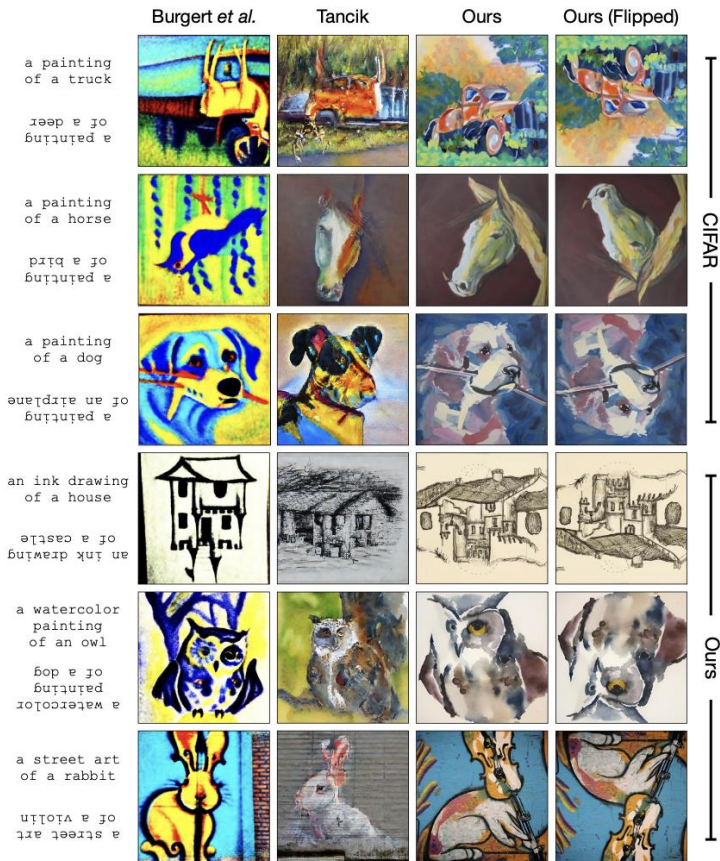
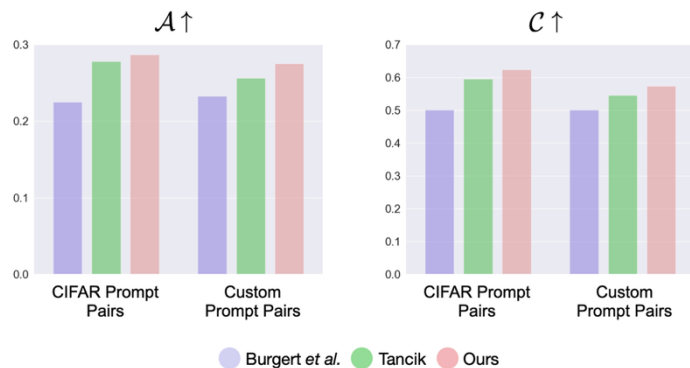


Figure 5. **Qualitative Comparisons.** We compare illusions generated by baselines to our illusions. We show examples from both our prompt dataset and the CIFAR prompt dataset.

4. Experiment

4.2 Ablations

1) Reduction Strategy

- Mean reduction $\epsilon_t^{avg} = \frac{1}{N} \sum_{i=1}^N v_i^{-1}(\hat{\epsilon}_i)$
- Alternating reduction

2) Negative Prompting

- Add negative prompts to suppress unwanted concepts

3) Guidance Scale

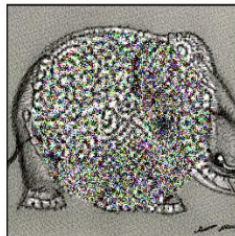
- Higher guidance scale have better illusion quality

Table 2. **Ablations.** We ablate negative prompting, reduction methods, and guidance scales on our dataset.

Ablation	$\mathcal{A} \uparrow$	$\mathcal{A}_{0.9} \uparrow$	$\mathcal{A}_{0.95} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{C}_{0.9} \uparrow$	$\mathcal{C}_{0.95} \uparrow$
Negative Prompting	0.24	0.27	0.276	0.576	0.659	0.683
No Negative Prompting	0.255	0.285	0.295	0.567	0.643	0.679
Alternating Reduction	0.252	0.286	0.292	0.560	0.639	0.664
Mean Reduction	0.255	0.285	0.295	0.567	0.643	0.679
$\gamma = 3.0$	0.239	0.271	0.285	0.537	0.610	0.629
$\gamma = 7.0$	0.255	0.285	0.295	0.567	0.643	0.679
$\gamma = 10.0$	0.259	0.290	0.297	0.576	0.664	0.702

4. Experiment

4.3 Failures



Failure: Correlated Noise

View: Rotate Inner Circle 45 Degrees (Bilinear)

a sketch of an elephant

a sketch of a mouse

1) Independent Synthesis

- The model rarely synthesizes prompts separately side-by-side rather than integrating them, due to center-bias in training data.

2) Noise Shift

- Linear but non-orthogonal transforms (e.g., White Balancing) alter Gaussian statistics, causing severe artifacting (spots).

3) Correlated Noise

- Rotations using standard bilinear sampling introduce correlations between adjacent pixels, breaking the assumption of independent noise and leading to divergent samples.

5. Limitation

1) Transformation Constraints: Cannot handle non-orthogonal deformations

- Scaling
- Affine Transformations
- Perspective Warping

2) Prompt Compatibility: Not all prompt pairs produce convincing illusions

- Semantically unrelated concepts
- Visually incompatible structures

3) Increasing Difficulty with More Views: As the number of view increases, multiple prompts compete for the same image structure, which leads to lower image quality and reduced semantic clarity.

6. Conclusion

1) Novel Zero-Shot Method

- First training-free framework for generating multi-view optical illusions
- Uses parallel denoising on pretrained text-to-image diffusion models

2) Theoretical Guarantee

- Valid view transformations must preserve Gaussian noise statistics
- Proven to be orthogonal transformations $A^T A = I$

3) Visual Anagrams (New Concept)

- Introduced a new class of optical illusions based on pixel permutations
- Extended to polymorphic jigsaw puzzles and arbitrary orthogonal views

4) Systematic Evaluation

- Demonstrated strong alignment and concealment performance
- Outperformed prior diffusion-based illusion generation methods