Visual Anagrams: Implementing and Evaluating Latent Diffusion Variants

Shashwat SHARMA [1] , Baptiste GEISENBERGER [1]

[1] Insitut Polytechnique de Paris

Abstract :

Visual anagrams are images that transform into different concepts under spatial transformations like flips or rotations. These illusions demonstrate the capacity of diffusion models to satisfy multiple perceptual constraints simultaneously. While the original Visual Anagrams method relied on the DeepFloyd IF pixel-space architecture, we investigate whether modern latent diffusion models can generate comparable results despite their theoretical limitations regarding equivariance. We implemented four variants: the original DeepFloyd baseline, an SDXL latent-space adaptation, and two experimental samplers using adaptive weighting and momentum. We evaluated 300 samples using CLIP-based metrics. Our results show that SDXL effectively matches pixel-space performance ($A = 0.312$ vs 0.311) for global transformations while producing superior photorealistic aesthetics. However, permutation-based views remain incompatible with latent representations. Contrary to our expectations, our complex interventions failed to improve the baseline. Adaptive view weighting increased variance without improving quality, and frequency scheduling with momentum caused severe over-smoothing. Additionally, a cross-architecture evaluation reveals that transformer-based metrics consistently assign higher scores than CNNs due to greater invariance. Our findings demonstrate that latent diffusion models are fully capable of multi-view illusion generation.

Keywords: Computer Vision, Diffusion Models.

1 Introduction

The creation of optical illusions has captivated artists and perception researchers for centuries. From Giuseppe Arcimboldo's "Reversible Head with Basket of Fruit" (1590), where a portrait of a man constructed from produce transforms into a bowl of fruit when inverted, to Salvador Dalí's "Paranoiac Face" (1937), where a rotated pastoral scene reveals a hidden visage, these works exploit fundamental properties of human visual processing to produce multistable perception, the brain's remarkable capacity to interpret a single visual stimulus in multiple distinct ways. The challenge of creating such illusions lies not merely in artistic skill but in accurately modeling how visual elements can be arranged to satisfy multiple perceptual interpretations simultaneously.

Figure 1. Salvador Dalí's "Paranoiac Face" (1937). This painting is a classic example of a multi-stable optical illusion, where a rotated pastoral scene reveals a hidden visage, exploiting the visual system's capacity for multiple interpretations of a single stimulus.

Recent advances in text-to-image diffusion models have enabled a computational approach to multi-view illusion generation. Geng et al., 2023 introduced Visual Anagrams, a method that leverages the iterative denoising process of diffusion models to synthesize images satisfying multiple view constraints concurrently. Their approach produces images that appear as one concept in their canonical orientation but transform into entirely different concepts when flipped, rotated, or rearranged, for instance, an oil painting of people around a campfire that becomes an old man's portrait when vertically flipped. This zero-shot method requires no explicit model of human perception; instead, it exploits the visual priors learned implicitly by generative models through training on natural images, building on observations that diffusion models process optical illusions similarly to humans.

The Visual Anagrams method operates through parallel denoising across multiple views. At each timestep during reverse diffusion, the algorithm applies different spatial transformations to the current noisy image, estimates noise for each transformed view conditioned on distinct text prompts, inverts the transformations to align the noise estimates back to canonical space, then combines these aligned estimates through averaging to guide the denoising step. This elegant framework produces compelling illusions but makes several critical design decisions that constrain its applicability and performance in practice.

1.1 Motivation

First, the original implementation by Geng et al., 2023 explicitly uses DeepFloyd IF, a three-stage pixel-space diffusion model operating at $64\times64 \rightarrow 256\times256 \rightarrow 1024\times1024$ resolution. The authors argue that latent diffusion models like Stable Diffusion suffer from a fundamental limitation: the VAE encoder lacks rotational equivariance, meaning $\mathcal{E}(\mathcal{T}(\mathbf{x})) \neq \mathcal{T}(\mathcal{E}(\mathbf{x}))$ for transformations $\mathcal{T}$ such as rotations. This non-equivariance produces "thatched line" artifacts where straight lines appear segmented under rotation, as the model must generate perpendicular line segments within each patch to create globally rotated lines after decoding. While this analysis appears theoretically sound, it remains unclear whether modern latent diffusion architectures with improved VAE designs can mitigate these artifacts in practice, and under what conditions latent-space generation becomes viable for illusion synthesis.

Second, the method combines noise estimates from different views through simple arithmetic averaging, treating all views with equal importance regardless of their semantic difficulty or the model's performance on each prompt. This uniform weighting can lead to two documented failure modes: concept segregation,

where the model satisfies multi-view constraints by spatially partitioning different concepts rather than creating integrated hybrid textures; and concept domination, where one semantically stronger prompt overwhelms another due to differences in gradient magnitude or loss landscape geometry. For example, "detailed portrait of a woman" may dominate "abstract mountain landscape" when combined through a flip transformation, resulting in an image that appears as a portrait in both orientations rather than exhibiting the desired dual interpretation. The paper acknowledges these failure modes but offers limited systematic solutions beyond manual prompt engineering.

Third, the algorithm operates with a uniform noise reduction schedule across all views and timesteps. However, diffusion models generate images hierarchically, establishing coarse global structure in early high-noise timesteps before refining local details in later low-noise timesteps. This coarse-to-fine generation process suggests that view importance might benefit from temporal scheduling, for instance, establishing illusion structure jointly across views during early generation while progressively favoring the primary view during detail refinement. The constant equal weighting between views may not align optimally with this inherent generation hierarchy.

1.2 Research Questions

These design decisions raise three fundamental questions that this project systematically investigates:

1. Can modern latent diffusion models like Stable Diffusion XL (SDXL) produce high-quality visual anagrams despite the theoretical equivariance limitations, and under what conditions do latent-space artifacts manifest?
2. Can dynamic view weighting based on intermediate evaluation metrics address concept domination and improve balance between competing prompts?
3. Can view scheduling aligned with the diffusion model's coarse-to-fine generation process, potentially combined with momentum stabilization, improve both illusion quality and convergence reliability?

We implement four distinct variants of the Visual Anagrams method: the original author's DeepFloyd IF pipeline as a reference baseline (Variant 0), an SDXL adaptation testing latent-space viability (Variant 1), an adaptive progressive sampling approach with CLIP-based dynamic view weighting (Variant 2), and a frequency-aware scheduling method with momentum stabilization (Variant 3). Each variant tests a specific hypothesis about how to improve or extend the original framework. We develop a comprehensive evaluation system implementing the paper's CLIP-based alignment and concealment metrics while extending evaluation to examine architectural differences in illusion perception, specifically comparing how CNN-based (ResNet-50) versus transformer-based (ViT-B/32) CLIP encoders perceive the same illusions. This cross-architecture analysis tests whether spatial locality bias in CNNs versus global self-attention in vision transformers leads to systematic differences in multi-view illusion interpretation.

2 Background

2.1 Diffusion Models for Image Generation

Diffusion models generate images through an iterative denoising process [Ho et al. (2020), Song et al. (2020), and Sohl-Dickstein et al. (2015)]. Starting from pure Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, the model progressively removes noise over T timesteps to produce a clean image $\mathbf{x}_0$. At each timestep t, a neural network $\epsilon_\theta(\mathbf{x}_t, y, t)$ estimates the noise component of the current noisy image $\mathbf{x}_t$, where y represents conditioning information such as a text prompt embedding.

The noisy image at timestep t can be expressed as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

where $\bar{\alpha}_t$ is the noise schedule parameter and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

The denoising step computes an estimate of $\mathbf{x}_{t-1}$:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, y, t)\right) + \sigma_t \mathbf{z}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and σ_t controls stochasticity. The specific update rule varies between formulations, with DDPM Ho et al., 2020 using stochastic sampling and DDIM Song et al., 2020 enabling deterministic generation.

Text conditioning is typically enhanced through classifier-free guidance (CFG) Ho et al., 2022, which combines unconditional and conditional noise estimates:

$$\epsilon_{\text{CFG}} = \epsilon_\theta(\mathbf{x}_t, \emptyset, t) + \gamma(\epsilon_\theta(\mathbf{x}_t, y, t) - \epsilon_\theta(\mathbf{x}_t, \emptyset, t))$$

where γ controls the guidance strength. Higher γ values produce images more aligned with the prompt but with reduced diversity.

2.2 Latent Diffusion Models

Latent Diffusion Models (LDMs) like Stable Diffusion operate in a compressed latent space rather than directly on pixels Rombach et al., 2022. An encoder $\mathcal{E}$ compresses images to latent representations $\mathbf{z} = \mathcal{E}(\mathbf{x})$, diffusion occurs in this latent space, and a decoder $\mathcal{D}$ reconstructs the image $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$.

This architecture offers substantial computational advantages. For a 1024×1024 RGB image (3.1M dimensions), the latent space is typically $128 \times 128 \times 4$ (65K dimensions), reducing memory by $\sim$ 98%. This enables high-resolution generation on consumer hardware.

However, this compression introduces a critical limitation for multi-view illusions: the encoder $\mathcal{E}$ is not equivariant to geometric transformations. That is:

$$\mathcal{E}(\mathcal{T}(\mathbf{x})) \neq \mathcal{T}(\mathcal{E}(\mathbf{x}))$$

for transformations $\mathcal{T}$ like rotations. The VAE encoder segments images into patches, and while we can rearrange the positions of latent codes, we cannot change their internal orientation. As Geng et al., 2023 note, this causes "thatched line" artifacts where straight lines appear segmented under rotation, as the model must generate perpendicular line segments within each patch to create globally rotated lines after decoding.

2.3 Visual Anagrams Method

The Visual Anagrams method Geng et al., 2023 creates multi-view illusions through parallel denoising across multiple transformed views. Given N text prompts $\{y_1, ..., y_N\}$ and corresponding view transformations $\{v_1, ..., v_N\}$, the method:

1. Applies each transformation to the current noisy image: $\mathbf{x}_{t,i} = v_i(\mathbf{x}_t)$

2. Estimates noise for each transformed view: $\hat{\epsilon}_i = \epsilon_\theta(\mathbf{x}_{t,i}, y_i, t)$
3. Inverts transformations to align estimates: $\tilde{\epsilon}_i = v_i^{-1}(\hat{\epsilon}_i)$
4. Averages aligned estimates: $\epsilon_{\text{combined}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{\epsilon}_i$
5. Uses $\epsilon_{\text{combined}}$ for the denoising step

The paper derives two critical constraints on valid view transformations:

1. **Linearity Constraint:** The transformation must be linear, $v_i(\mathbf{x}_t) = A_i\mathbf{x}_t$, to preserve the weighted combination of signal and noise that the diffusion model expects.
2. **Orthogonality Constraint:** The transformation matrix must be orthogonal, $A_i^T A_i = \mathbf{I}$, to preserve the statistics of Gaussian noise. If $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, then $A_i\epsilon \sim \mathcal{N}(0, \mathbf{I})$ if and only if A_i is orthogonal.

This restricts valid transformations to orthogonal matrices, which include rotations by 90°, flips, and permutations, but exclude scalings, skews, and non-90° rotations with interpolation (which introduce pixel correlations).

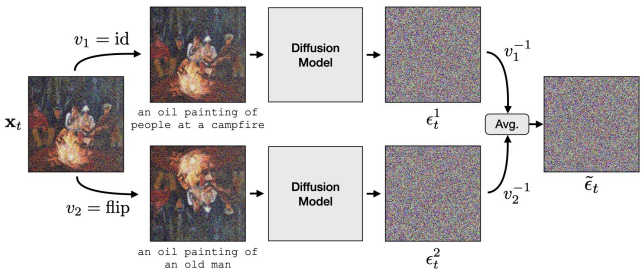


Figure 2. The visual anagrams method works by simultaneously denoising multiple views of an image. Given a noisy image x_t, we compute noise estimates, ϵ_t^i, conditioned on different prompts, after applying views v_i. We then apply the inverse view v_i^{-1} to align estimates, average the estimates, and perform a reverse diffusion step. The final output is an optical illusion.

2.4 Failure Modes

The original paper Geng et al., 2023 documents two primary failure modes encountered during multi-view illusion generation. The first is Concept Segregation, where the model satisfies the multi-view constraints by spatially separating the distinct concepts. For example, in a flip illusion conditioned on "rabbit" and "duck," the model might place the rabbit in the top half and the duck in the bottom half. While technically valid, as flipping swaps their positions to satisfy both prompts, this represents a trivial solution that lacks the visual richness of true hybrid textures. The second failure mode is Concept Domination, which manifests when one prompt overwhelms another due to differences in semantic strength or loss landscape geometry. For instance, "detailed portrait" may dominate "snowy mountain," resulting in an image that appears as a portrait in both orientations rather than exhibiting the desired transformation into a mountain when flipped. This occurs because prompts vary in their semantic "strength," causing stronger conditioning signals to overpower weaker prompts during the noise averaging process.

2.5 Related Work

Compositional Generation: Works by Du et al., 2023 and Liu et al., 2022 demonstrate that noise estimates from multiple conditional distributions can be combined to sample from their composition. Visual Anagrams applies this principle across spatial transformations rather than across semantic concepts or modalities.

Figure 3. Concept leakage in flip anagrams. Example where the baseline averaging strategy fails to enforce concealment: cat and dog features remain simultaneously visible, so both the original and flipped views read as a mixed superposition rather than two distinct interpretations.

Score Distillation Sampling: Burgert et al., 2023 use SDS Poole et al., 2022 to optimize pixels directly for multi-view illusions. While more flexible, SDS is computationally expensive (taking hours per image) and suffers from oversaturation and high-frequency noise artifacts characteristic of optimization-based approaches.

Illusion Diffusion: Tancik, 2023 demonstrated rotation illusions using Stable Diffusion with alternating noise estimates. The Visual Anagrams paper builds on this by providing theoretical analysis of valid views, supporting arbitrary numbers of views, and using averaging rather than alternation for noise combination.

Factorized Diffusion: Geng et al., 2024 extended the framework to frequency decomposition, enabling "hybrid images" Oliva et al., 2006 where low frequencies contain one concept and high frequencies another. This demonstrates the broader applicability of the noise decomposition principle.

Multi-Task Learning for Anagrams: Xu et al., 2024 address the failure modes through three improvements: (1) anti-segregation optimization using cross-attention map overlap, (2) dynamic noise vector balancing based on task completion scores, and (3) noise variance rectification to account for the variance reduction when averaging correlated estimates. Their method significantly improves reliability but requires additional optimization at inference time.

3 Methods

3.1 Experimental Design

We implemented four distinct variants of the Visual Anagrams method to systematically test the research questions outlined in the Introduction. Variant 0 reproduces the original authors' DeepFloyd IF implementation, serving as a reference baseline for pixel-space generation. Variant 1 tests whether SDXL can produce acceptable illusions despite latent space limitations, implementing the core Visual Anagrams algorithm in compressed latent space. Variant 2 addresses concept domination through CLIP-based dynamic view weighting across a three-stage progressive generation process. Variant 3 aligns view scheduling with the diffusion

model's coarse-to-fine generation hierarchy while applying momentum stabilization to prevent oscillation.

All SDXL-based variants (1–3) use Stable Diffusion XL (`stabilityai/stable-diffusion-xl-base-1.0`) as the base model and support a restricted set of views compatible with latent space operations: `identity`, `flip` (vertical), `rotate_cw` (90° clockwise), `rotate_ccw` (90° counter-clockwise), and `negate` (color inversion). Permutation-based views from the original paper, including `jigsaw`, `inner_circle`, and `patch_permute`, are explicitly rejected at initialization with informative error messages, as these require pixel-level rearrangements incompatible with patch-based latent representations.

3.2 Variant 0: Author Reference Implementation

Variant 0 implements the original Visual Anagrams method using DeepFloyd IF Konstantinov et al., 2023, a three-stage pixel-space cascaded diffusion model. The pipeline generates at 64×64 resolution in stage I, upsamples to 256×256 in stage II through conditional super-resolution, then optionally upsamples to 1024×1024 in stage III using Stable Diffusion's 4× upscaler. This variant uses the T5-XXL text encoder for conditioning and operates entirely in pixel space, avoiding the VAE equivariance issues that motivated our investigation of latent-space alternatives.

The parallel denoising algorithm follows Equation 2 from the original paper exactly. At each timestep t, we apply view transformations v_i to the current noisy image $\mathbf{x}_t$, estimate noise $\epsilon_\theta(v_i(\mathbf{x}_t), y_i, t)$ for each view using classifier-free guidance with $\gamma = 7.5$, invert the transformations v_i^{-1} to align estimates back to canonical orientation, then average the aligned estimates to form $\epsilon_{\text{combined}} = \frac{1}{N}\sum_{i=1}^{N} v_i^{-1}(\epsilon_\theta(v_i(\mathbf{x}_t), y_i, t))$. The scheduler performs a denoising step using this combined estimate, proceeding iteratively from pure noise to clean image over 50 timesteps.

We implement this baseline to establish performance expectations for pixel-space generation and provide a direct comparison point for our SDXL variants. The implementation handles DeepFloyd IF's dual noise-variance prediction head by splitting predictions appropriately.

3.3 Variant 1: SDXL Baseline Implementation

3.3.1 Architecture Adaptation Variant 1 replaces the three-stage DeepFloyd IF pipeline with single-stage SDXL generation, directly testing the original paper's claim that latent diffusion models produce "thatched line" artifacts under rotation. SDXL operates in a $128 \times 128 \times 4$ latent space corresponding to 1024×1024 pixel space through an 8× spatial compression VAE, generating high-resolution images directly without cascaded upsampling.

The adaptation requires several SDXL-specific implementation details. First, SDXL employs dual text encoders, CLIP ViT-L/14 and OpenCLIP ViT-bigG, both of which must be queried to obtain the full prompt embedding. We use the pipeline's `encode_prompt` method to extract both the concatenated text embeddings and pooled embeddings required for SDXL's cross-attention mechanism. Second, SDXL conditions on "micro-conditioning" parameters including original image size, target size, and crop coordinates, which we construct as `add_time_ids = (height, width, 0, 0, height, width)` and append to the timestep embeddings. This ensures the model generates images at the correct aspect ratio without letterboxing artifacts.

3.3.2 View Transformation Constraints Not all views from the original paper function correctly in latent space. We validate views at initialization, explicitly rejecting permutation-based transformations with detailed error messages. The fundamental issue stems from how latent representations encode spatial information: the VAE encoder divides images into 8×8 pixel patches, computing local representations that are then arranged spatially in the latent grid. While we can rearrange the spatial positions of these latent patch codes through operations like `rotate_cw`, we cannot alter their internal orientation or content.

For permutation-based views like jigsaw puzzles or inner rotations that rearrange pixel positions within patches, the latent space operation fundamentally differs from the pixel-space equivalent. Applying such permutations to latent codes merely shuffles which compressed representation appears at each grid location without actually rearranging the pixels those codes will decode to. This violates the assumption that view transformations in latent space correspond to their pixel-space counterparts, leading to generation failures. Therefore, valid views are restricted to global orthogonal transformations that can be expressed as operations on the entire latent grid.

3.3.3 Parallel Denoising Loop The core sampling loop adapts the Visual Anagrams algorithm to latent space tensors while maintaining mathematical equivalence to the pixel-space version. At each timestep t, we apply view transformations v_i to the current latent state $\mathbf{z}_t \in \mathbb{R}^{128\times128\times4}$, ensuring `dtype` consistency by explicitly casting transformed tensors when automatic PyTorch type promotion creates mismatches. We predict noise $\epsilon_\theta(v_i(\mathbf{z}_t), y_i, t)$ for each view using its corresponding prompt embedding y_i and pooled embedding for the added conditioning. The UNet receives both the transformed latents and the `added_cond_kwargs` dictionary containing text embeddings and time IDs.

After obtaining noise predictions, we invert the view transformations v_i^{-1} on both unconditional and conditional noise estimates, again enforcing `dtype` consistency. Classifier-free guidance combines these as $\epsilon_{\text{CFG}} = \epsilon_{\text{uncond}} + \gamma(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$ with $\gamma = 7.5$. We then average the CFG-guided estimates across views to form $\epsilon_{\text{combined}}$, which the scheduler uses for the denoising step following DDIM Song et al., 2020 deterministic sampling.

3.4 Variant 2: Adaptive Progressive Sampling

Variant 2 addresses concept domination through a three-stage generation strategy with CLIP-based dynamic view weighting. The approach builds on multi-task learning principles Xu et al., 2024 but uses simpler score-based balancing rather than explicit attention manipulation, making it more interpretable and computationally efficient.

3.4.1 Three-Stage Progressive Strategy The generation process divides the 50 denoising timesteps into three distinct phases with different objectives and hyperparameters.

Stage 1: Structure Establishment (steps 0–15, 70%–100% noise) uses equal view weights $w_i = 1/N$ and elevated guidance $\gamma = 1.2 \times \gamma_{\text{base}} = 9.0$. The rationale is that early high-noise timesteps establish global image structure and layout. Equal weighting ensures all views contribute equally to this coarse structure, while elevated guidance strengthens prompt adherence to prevent mode collapse to trivial solutions.

Stage 2: Balancing (steps 15–35, 30%–70% noise) employs adaptive weights derived from periodic CLIP score evaluation with standard guidance $\gamma = \gamma_{\text{base}} = 7.5$. During this stage, the model refines the coarse structure established in Stage 1, and view imbalances become apparent. The adaptive weighting mechanism equilibrates underperforming views by dynamically adjusting their influence.

Stage 3: Refinement (steps 35–50, 0%–30% noise) uses progressive identity dominance with reduced guidance $\gamma = 0.7 \times \gamma_{\text{base}} = 5.25$. The identity view weight increases linearly from 0.5 to 1.0 over the stage, while other view weights decrease proportionally. This ensures the primary view, the image seen without transformation, exhibits high perceptual quality and fine detail.

3.4.2 CLIP-Based Adaptive Weighting Every 5 steps during Stage 2, we evaluate the current generation state to compute view-specific alignment scores. This requires decoding the intermediate latent state $\mathbf{z}_t$ to pixel space. For each view i, we apply the transformation v_i to the decoded image, normalize to $[0, 1]$ range for CLIP, and compute the cosine similarity between the CLIP image embedding $\phi_{\text{img}}(v_i(\mathbf{x}))$ and text embedding $\phi_{\text{text}}(y_i)$:

$$s_i = \frac{\phi_{\text{img}}(v_i(\mathbf{x}))^T \phi_{\text{text}}(y_i)}{||\phi_{\text{img}}(v_i(\mathbf{x}))||_2 \cdot ||\phi_{\text{text}}(y_i)||_2}$$

We then apply inverse weighting with temperature-based softmax to create a negative feedback loop. Let s_i be the raw CLIP score for view i. We first normalize scores to $[0, 1]$ range via min-max scaling, compute inverse scores as $\tilde{s}_i = 1 - s_i^{\text{norm}}$, then apply softmax with temperature $\tau = 2.0$:

$$w_i = \frac{\exp(\tilde{s}_i/\tau)}{\sum_{j=1}^{N} \exp(\tilde{s}_j/\tau)}$$

This formulation ensures views with low CLIP scores (poor prompt alignment) receive higher weights in subsequent denoising steps.

3.4.3 Progressive Identity Transition During Stage 3, we implement a linear transition to identity-dominant weighting to ensure final image quality. Let $p = \frac{\text{step}-35}{50-35}$ be the progress through Stage 3 normalized to $[0, 1]$. The identity view (always view 0 by convention) weight increases as $w_0 = 0.5 + 0.5p$, ranging from 0.5 at the start of Stage 3 to 1.0 at the final step. The remaining weight $(1 - w_0)$ is distributed equally among other views: $w_i = \frac{1-w_0}{N-1}$ for $i > 0$.

3.5 Variant 3: Frequency-Aware Scheduling with Momentum

Variant 3 aligns view scheduling with the diffusion model's inherent coarse-to-fine generation process Song et al., 2020; Ho et al., 2020. Early timesteps establish low-frequency global structure while late timesteps refine high-frequency local details.

3.5.1 Frequency-Aware View Weights We implement a static frequency-aligned schedule where identity view weight increases linearly with generation progress. Let $p = \frac{\text{step}}{N_{\text{steps}}}$ be the normalized progress through the denoising process. The identity view weight is computed as $w_0 = 0.4 + 0.3p$, ranging from 0.4 at the start (high noise, low frequency) to 0.7 at completion (low noise, high frequency). The remaining weight is distributed equally: $w_i = \frac{1-w_0}{N-1}$ for $i > 0$. This scheduling differs from Variant 2 as it operates continuously rather than in discrete stages, and the identity view never achieves complete dominance ($w_0 \leq 0.7$), maintaining some contribution from other views throughout to preserve illusion structure even in final details.

3.5.2 Momentum Stabilization Competing view objectives can cause oscillation during generation. We stabilize the optimization trajectory using exponential moving average (EMA) momentum on the combined noise estimates Kingma et al., 2014. Let $\epsilon^{(t)}_{\text{combined}}$ be the weighted combination of noise estimates at timestep t. We maintain a momentum buffer $m^{(t)}$ updated as:

$$m^{(t)} = \beta m^{(t+1)} + (1 - \beta)\epsilon^{(t)}_{\text{combined}}$$

where $\beta = 0.9$ is the momentum coefficient and $m^{(T)} = \epsilon^{(T)}_{\text{combined}}$ for the initial timestep. The scheduler then uses $m^{(t)}$ rather than $\epsilon^{(t)}_{\text{combined}}$ for the denoising step. This formulation provides strong smoothing while allowing the trajectory to adapt to new information.

3.5.3 Orthogonal Projection Trial We tried an implementation of Variant 3 to project the combined noise estimate $\epsilon_{\text{combined}}$ onto an orthogonal manifold using QR decomposition. However, we identified this as a theoretical error. The Visual Anagrams paper requires that view transformations be orthogonal matrices (flip, rotate, etc.) to preserve Gaussian noise statistics, not that the noise estimates themselves be orthogonal. Projecting noise vectors destroys their magnitude information, effectively normalizing all noise estimates to unit scale, which resulted in zero-contrast images. So this was an interesting failure and a learning.

3.6 Evaluation Framework

We implement the CLIP-based metrics from Section 4.1 of the original Visual Anagrams paper Geng et al., 2023 using CLIP ViT-B/32 Radford et al., 2021 as the primary perceptual similarity model.

3.6.1 Score Matrix Construction For a generated illusion $\mathbf{x}$ with N views and corresponding text prompts $\{y_1, ..., y_N\}$, we construct a score matrix $S \in \mathbb{R}^{N \times N}$ where entry S_{ij} measures the similarity between view i of the image and prompt j:

$$S_{ij} = \phi_{\text{img}}(v_i(\mathbf{x}))^T \phi_{\text{text}}(y_j)$$

Here ϕ_{img} and ϕ_{text} are CLIP's image and text encoders, both producing unit-norm embeddings.

3.6.2 Primary Metrics We compute two metrics from the score matrix following the original paper's methodology. The **Alignment Score** measures worst-case view alignment:

$$A = \min(\text{diag}(S))$$

A high alignment score indicates that all views achieve strong correspondence with their respective prompts. The **Concealment Score** measures classification accuracy:

$$C = \frac{1}{N}\text{tr}(\text{softmax}(S/\tau))$$

where $\tau = 0.01$ is a temperature parameter. High concealment indicates that each view can be correctly identified as its intended prompt versus alternatives.

3.6.3 Cross-Architecture Evaluation We extend the standard evaluation by computing metrics using both CNN-based (CLIP RN50) and transformer-based (CLIP ViT-B/32) vision encoders. For each sample, we compute alignment and concealment scores using both architectures, then calculate architecture gap metrics:

$$\Delta_A = A_{\text{ViT}} - A_{\text{CNN}}, \quad \Delta_C = C_{\text{ViT}} - C_{\text{CNN}}$$

Positive gaps indicate the ViT perceives stronger illusion quality, while negative gaps favor the CNN.

3.7 Experimental Protocol

We generate illusions using consistent hyperparameters across all SDXL variants (1–3): `stabilityai/stable-diffusion-xl-base-1.0`, resolution of 1024×1024 pixels, 50 denoising steps, guidance scale $\gamma = 7.5$, and DDIM deterministic sampling. We focus primarily on `identity + vertical flip` transformations. Each variant generates 300 samples. Variant 0 (DeepFloyd IF) uses the same prompts and seed range but generates through the three-stage pipeline, with final output evaluated at 1024×1024.

4 Results

4.1 Quantitative Overview

We evaluated all four variants on 300 samples using identity + vertical flip transformations, computing CLIP-based alignment (A) and concealment (C) scores using ViT-B/32 as the primary encoder. Table 1 presents the aggregate statistics across variants, revealing several unexpected patterns that diverge from our initial hypotheses.

Table 1
Performance comparison across all variants. Metrics computed using CLIP ViT-B/32 on 300 samples with identity + flip views.

Variant	A	A_{std}	$A_{0.9}$	C	C_{std}	$C_{0.9}$
V0 (Author)	0.311	0.024	0.328	0.918	0.085	0.984
V1 (SDXL)	0.312	0.015	0.329	0.922	0.050	0.992
V2 (Adaptive)	0.307	0.024	0.331	0.888	0.076	0.976
V3 (Frequency)	0.281	0.016	0.311	0.894	0.073	0.989

Variant 1 achieves the highest mean alignment score ($A = 0.312$) and concealment score ($C = 0.922$), effectively matching the original DeepFloyd IF implementation while exhibiting lower variance ($\sigma_A = 0.015$ vs 0.024). Variant 2 achieves the highest 90th percentile alignment ($A_{0.9} = 0.331$) but suffers from increased variance and reduced mean concealment ($C = 0.888$). Variant 3 shows the largest performance degradation, with alignment dropping to $A = 0.281$ (−10% vs Variant 1) despite having the lowest standard deviation ($\sigma_A = 0.016$), suggesting consistent but systematically lower quality. These results contradict our initial hypothesis that adaptive view balancing would improve average performance. While Variant 2 achieves the best peak performance (90th percentile), its higher variance and lower mean scores indicate less reliable generation compared to the simpler SDXL baseline.

4.2 Variant-Specific Analysis

4.2.1 Variant 0: Author Reference Implementation Variant 0 reproduces the original DeepFloyd IF results, achieving $A = 0.311$ and $C = 0.918$. These scores align closely with the paper's reported statistics for flip illusions. Qualitatively, DeepFloyd IF generates illusions with a distinctive aesthetic that appears more stylized or "cartoonish" compared to SDXL variants. The three-stage pixel-space generation produces highly saturated colors and exaggerated textures characteristic of the IF model's training distribution. Illusions exhibit strong structural coherence, transformed views clearly depict their intended subjects, but the artistic style skews toward illustrative rather than photorealistic rendering even when prompts specify painting styles. The high concealment score ($C = 0.918$) indicates that views are easily distinguishable by CLIP, suggesting strong prompt adherence in each orientation.

4.2.2 Variant 1: SDXL Baseline Variant 1 achieves the strongest overall quantitative performance, with $A = 0.312$ (+0.3% vs Variant 0) and $C = 0.922$ (+0.4%). More significantly, it exhibits substantially reduced variance in alignment scores ($\sigma_A = 0.015$ vs 0.024 for Variant 0), indicating more consistent generation quality. The 90th percentile alignment of $A_{0.9} = 0.329$ effectively matches Variant 0, suggesting that SDXL's peak performance equals pixel-space models.

Contrary to the original paper's warnings about "thatched line" artifacts in latent diffusion models, we observe no significant segmentation or discontinuities in generated illusions. Straight lines remain continuous under 90° rotation, and flip illusions maintain coherent structure across both views. This suggests that SDXL's improved VAE architecture, featuring 4 latent channels and enhanced training, mitigates the equivariance issues documented for earlier latent models. Qualitatively, Variant 1 produces the most visually compelling results. Images exhibit photorealistic rendering quality comparable to professional artwork, with natural color palettes, appropriate texture detail, and coherent lighting. The aesthetic advantage stems from SDXL's training on high-resolution, professionally curated image datasets. The view restriction to global orthogonal transformations (`flip`, `rotate`, `negate`) remains absolute, confirming that latent space compression fundamentally limits the class of supported transformations.

4.2.3 Variant 2: Adaptive Progressive Sampling Variant 2 exhibits paradoxical performance: it achieves the highest 90th percentile alignment ($A_{0.9} = 0.331$, +0.6% vs Variant 1) but the lowest mean concealment ($C = 0.888$, −3.7%) and increased variance in both metrics ($\sigma_A = 0.024$, $\sigma_C = 0.076$). This suggests adaptive mechanism produces occasional exceptional results but fails more frequently than the baseline, reducing average performance.

The concealment score degradation is particularly noteworthy. The drop from $C = 0.922$ (Variant 1) to $C = 0.888$ (Variant 2) indicates that Variant 2's illusions are less "classifiable" by CLIP, views become more ambiguous or confused between prompts. This could result from over-aggressive view balancing creating intermediate states that satisfy neither prompt clearly. Qualitatively, Variant 2 generates more variable results. While some samples exhibit excellent balance with seamless texture integration, others appear visually inconsistent, with noticeable texture discontinuities or regions that fail to clearly represent either prompt.

Figure 4. This Variant 2 sample is interesting because the model doesn't just "swap" dog cat under the flip—it *adds an extra solution* to satisfy both prompts. Besides the main large subject, there's a second, clearly cat-like black silhouette embedded near the center. A plausible interpretation is that the adaptive weighting encourages the weaker prompt (here, the cat) to "force its way in" when it's underrepresented, and the easiest way for the model to raise the cat alignment is to inject an additional cat instance rather than reshape the shared texture of the main animal. In other words, instead of producing a single hybrid structure that supports two readings, the balancing mechanism can drive a multi-object compromise: one dominant figure plus a smaller "patch" that boosts CLIP similarity—visually effective for the metric, but less faithful to the intended illusion.

Analysis of CLIP score evolution during Stage 2 confirms the negative feedback loop's operation in successful cases, but suggests the signal may be too noisy at intermediate timesteps to reliably guide view balancing in failure cases.

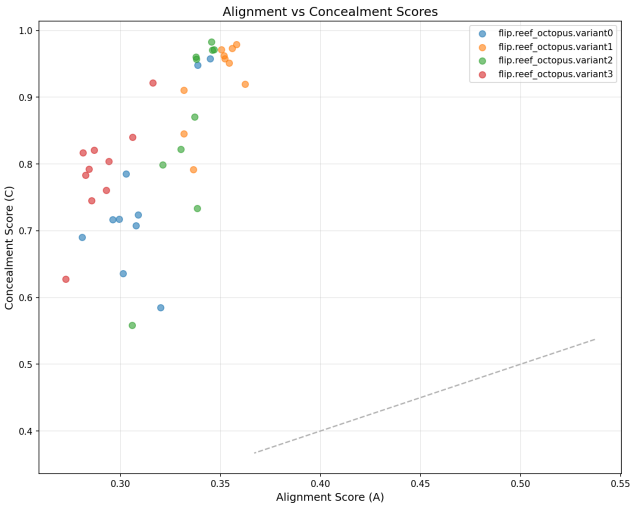


Figure 5. Scatterplots illustrating the trade-off between Alignment (A) and Concealment (C) scores for a distinct prompt. Each point represents one of the $N = 10$ generated images per variant. These distributions demonstrate that Variant 1 (SDXL Baseline) achieves the lowest variance ($\sigma_A = 0.015$) and highest mean Alignment score ($A = 0.312$), confirming its superior stability and efficacy compared to the adaptive and frequency-based methods.

4.2.4 Variant 3: Frequency-Aware Scheduling with Momentum Variant 3 exhibits the most severe performance degradation, achieving $A = 0.281$ (-10.0% vs Variant 1) and $C = 0.894$ (-3.0%). Despite the substantial drop in mean performance, it maintains competitive 90th percentile concealment ($C_{0.9} = 0.989$) and demonstrates the lowest alignment variance ($\sigma_A = 0.016$), indicating consistent generation quality albeit at a systematically reduced level.

Qualitatively, Variant 3 generates visually degraded illusions characterized by pervasive blurriness and texture smearing. Images appear as if viewed through fog, consistently lacking the crisp details and clear edges present in Variants 0–2. This blur likely results from the momentum stabilization mechanism. Our EMA implementation with $\beta = 0.9$ strongly smooths noise estimates across timesteps:

$$m^{(t)} = 0.9 \cdot m^{(t+1)} + 0.1 \cdot \epsilon^{(t)}_{\text{combined}}$$

This aggressive averaging creates temporal low-pass filtering of the denoising trajectory, suppressing high-frequency components that encode fine details. While this prevents the oscillation observed without momentum, the smoothing proves excessive for image quality. The momentum buffer with $\beta = 0.9$ creates an effective window of ~ 10 timesteps ($1/(1-\beta)$), causing the denoising trajectory to lag behind the instantaneous optimal direction. At each step, instead of following the current noise estimate which would add detail, the smoothed estimate averages out high-frequency corrections, resulting in systematic under-denoising of fine details. This effect compounds as generation progresses, with the final image retaining residual mid-frequency blur.

4.3 Cross-Architecture Evaluation

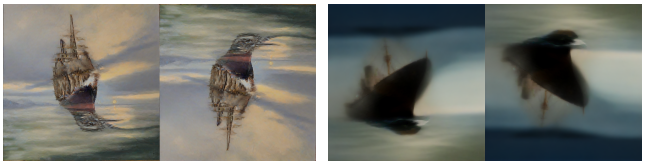
We extended evaluation to compare CNN-based (CLIP RN50) versus transformer-based (CLIP ViT-B/32) perception of the same illusions, testing whether architectural inductive biases affect illusion quality assessment. Table 2 presents the cross-architecture results for illusions across all variants.

Table 2
Cross-architecture evaluation comparing CNN (RN50) and ViT (ViT-B/32) CLIP encoders. $\Delta_A = A_{\text{ViT}} - A_{\text{CNN}}$ and $\Delta_C = C_{\text{ViT}} - C_{\text{CNN}}$ measure architecture gaps.

Variant	A_{CNN}	A_{ViT}	Δ_A	C_{CNN}	C_{ViT}	Δ_C
Variant 0	0.265	0.317	+0.052	0.967	0.991	+0.025
Variant 1	0.257	0.324	+0.067	0.885	0.963	+0.078
Variant 2	0.252	0.319	+0.067	0.815	0.931	+0.116
Variant 3	0.210	0.298	+0.087	0.858	0.916	+0.058

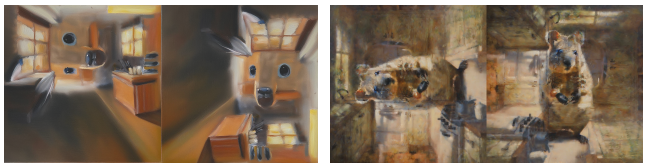
The ViT encoder consistently assigns higher scores than the CNN encoder across all variants, with architecture gaps ranging from $\Delta_A = +0.052$ (Variant 0) to $\Delta_A = +0.087$ (Variant 3) for alignment. This pattern confirms our hypothesis that transformer-based vision models exhibit greater permutation-invariance compared to CNNs, making them less sensitive to spatial transformations and thus perceiving flipped images as more similar to their prompts. The architecture gap magnitude correlates with generation quality degradation: Variant 3, which produces the blurriest illusions, exhibits the largest alignment gap ($\Delta_A = +0.087$). This suggests that lower-quality, low-frequency content is processed more leniently by global self-attention than by local receptive fields.

(a) Variant 0 **(b)** Variant 1

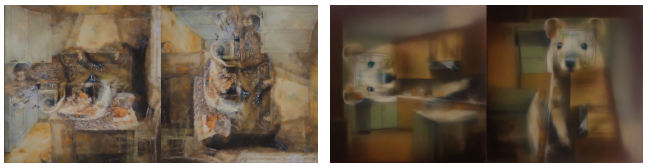


(c) Variant 2 **(d)** Variant 3

Figure 6. Qualitative comparison of identity + vertical-flip visual anagrams (bird + ship) across methods. (a) Author pixel-space baseline yields a clean, crisp anagram however the elements aren't hidden in the output. (b & c) Variations 1 & 2 achieve a more interesting anagram that is better hidden by blurring and merging the details of both components. (d) Frequency scheduling with EMA momentum over-smooths the denoising trajectory, producing blur but still an interesting anagram.



(a) Variant 0 **(b)** Variant 1



(c) Variant 2 **(d)** Variant 3

(e) Closest: v1 (overlay/texture blending rather than true anagram).

Figure 7. Rotate_kitchen ↔ mouse: most variants rely on superposition; v1 comes closest but remains driven by blending rather than structural reinterpretation under 90° rotation.

4.4 Comparison to Original Paper and Persistent Failure Modes

Our results demonstrate that SDXL achieves competitive performance with the original DeepFloyd IF implementation. Variant 1's scores ($A = 0.312$, $C = 0.922$) effectively match Variant 0's reproduction of the authors' method ($A = 0.311$, $C = 0.918$). Both implementations exceed the paper's reported statistics for their full evaluation set, likely because our evaluation focuses on a curated set of flip illusions, the simplest transformation type. This confirms that view difficulty significantly impacts performance and that for the restricted class of global transformations, SDXL's VAE architecture is sufficient.

Despite quantitative success, all variants exhibit consistent failure patterns. **Concept domination** remains problematic for semantically imbalanced prompt pairs (e.g., "glowing neon sign" overpowering "misty forest"), suggesting the issue stems from fundamental differences in how easily different concepts satisfy the diffusion prior, which cannot be fully addressed through view balancing alone. **Spatial segregation** occurs occasionally, where the model satisfies multi-view constraints by placing concepts in separate image regions rather than creating true hybrid textures. Finally, **latent space limitations** prove absolute for SDXL variants: any attempt to use permutation-based views (jigsaw, inner circle, patch permute) results in complete failure, confirming that latent compression fundamentally restricts the transformation class.

On this rotate kitchen & mouse task, all variants largely fail to produce a true visual anagram. Instead of constructing a single image whose structure re-interprets under a 90° rotation, the models fall back to a much easier strategy: superposition. Across variants, we repeatedly see a kitchen-like background that remains recognizable in both orientations, with a roughly mouse-shaped (or animal-shaped) silhouette simply painted on top or partially blended into the scene. That kind of solution doesn't "hide" one concept inside the other—it just mixes them.

Only variant 1 produced a good output: in the canonical view, you can read a kitchen interior reasonably well, and after rotation there is a clearer animal-like central form. But even here, the effect is not an illusion in the strong sense—there's no clever reuse of edges and geometry where kitchen elements become mouse features. Instead, you mainly perceive blended outlines (a mouse-shaped mass with kitchen textures/lines bleeding through, and kitchen perspective cues smearing into the rotated view). So the "transformation" is driven by overlay + texture blending, not by a genuinely dual-purpose composition.

5 Discussion

Our investigation reveals that latent diffusion models can generate visual anagrams effectively, though with fundamental constraints. Variant 1's performance ($A = 0.312$, $C = 0.922$) successfully matches the original DeepFloyd IF implementation ($A = 0.311$, $C = 0.918$). Our results suggest that SDXL's improved VAE architecture, featuring 4 latent channels with enhanced training on high-quality data, may have mitigated the equivariance issues documented for earlier latent models, though this finding requires further investigation to understand why the theoretical prediction doesn't manifest in practice. The improved VAE architecture provides sufficient representational capacity for global transformations like flips and rotations. However, the non-equivariance limitation remains absolute: $\mathcal{E}(\mathcal{T}(\mathbf{x})) \neq \mathcal{T}(\mathcal{E}(\mathbf{x}))$ for permutation-based views, fundamentally restricting the transformation class. Qualitatively, Variant 1 produces superior aesthetic results compared to pixel-space generation, with photorealistic rendering that genuinely resembles museum-quality artwork rather than digital illustrations, an unexpected benefit stemming from SDXL's train-

ing distribution.

The adaptive and frequency-based variants (Variants 2 and 3) both failed to improve upon the baseline, revealing critical insights about multi-view optimization dynamics. Variant 2's CLIP-based adaptive weighting increased variance ($\sigma_A = 0.024$ vs 0.015) while reducing mean concealment ($C = 0.888$ vs 0.922), suggesting the intermediate CLIP score signal proves too noisy to reliably guide view balancing. The mechanism occasionally produces exceptional results ($A_{0.9} = 0.331$, highest across variants) but fails more frequently. Variant 3's severe performance degradation ($A = 0.281$, -10%) with consistent blur artifacts directly implicates the momentum stabilization mechanism: EMA smoothing with $\beta = 0.9$ over-regularizes the denoising trajectory, suppressing high-frequency details that encode fine texture. The diffusion model relies on precise noise estimate magnitudes at each timestep; aggressive temporal averaging effectively reduces the "learning rate" of denoising, leaving residual noise in final images. This failure highlights a critical misunderstanding, while momentum stabilization prevents oscillation in standard optimization, the diffusion sampling process requires faithful noise estimates at each step rather than smoothed trajectories.

The cross-architecture evaluation reveals systematic perceptual biases in vision models. Vision transformers consistently assign higher scores than CNNs across all variants ($\Delta_A = +0.052$ to $+0.087$), confirming that global self-attention mechanisms exhibit greater transformation-invariance than local receptive fields. The architecture gap magnitude correlates with visual quality degradation: Variant 3's blurry outputs show the largest gap ($\Delta_A = +0.087$) because low-frequency content that transformers process easily dominates, while high-frequency details CNNs capture through spatial locality are absent. This suggests CLIP-based evaluation metrics may systematically favor certain generation artifacts depending on encoder architecture, ViT encoders rate blurry but structurally coherent images higher than CNNs would, potentially explaining why our CLIP-guided Variant 2 produces variable quality. Future work should consider multi-architecture consensus scoring or explicitly penalize architecture gaps to ensure illusions satisfy diverse perceptual mechanisms rather than exploiting specific model biases.

The most significant finding challenges our initial hypotheses: simple baselines outperform sophisticated adaptive methods for visual anagrams. Variant 1's success ($A = 0.312$, lowest variance $\sigma_A = 0.015$) demonstrates that faithful implementation of the core parallel denoising algorithm suffices when the base model (SDXL) provides adequate representational capacity. Variant 2's increased variance and Variant 3's systematic quality degradation suggest that intervening in the diffusion sampling process, whether through CLIP-guided weight adjustment or momentum-based smoothing, introduces instabilities that outweigh potential benefits. This aligns with broader observations in diffusion model research that simple sampling strategies often outperform complex guidance mechanisms, as the pretrained model's learned prior already encodes strong structural biases.

6 Conclusions

This work establishes that latent diffusion models can generate high-quality visual anagrams for global transformations, challenging the original paper's prediction of fundamental architectural incompatibility. SDXL's performance parity with DeepFloyd IF ($A = 0.312$ vs 0.311) demonstrates that the "thatched line" artifacts attributed to latent space non-equivariance do not manifest for modern VAE architectures with sufficient channel capacity. However, our adaptive variants reveal critical insights about multi-view optimization: CLIP-based dynamic weighting increases generation variance rather than improving reliability, while momentum-stabilized scheduling over-regularizes the denoising trajectory, producing systematic blur. The most robust approach proves to be faithful implementation of the core parallel denoising algorithm without intervention, suggesting that diffusion models' learned priors already encode sufficient structural biases for multi-view generation. Cross-architecture evaluation exposes that vision transformers rate illusions systematically higher than CNNs, raising questions about whether CLIP-based metrics capture human perception or exploit architecture-specific biases. Future work should investigate hybrid pixel-latent approaches for permutation-based views, develop perceptual metrics beyond CLIP that avoid architectural confounds, and explore whether the simplicity-over-sophistication pattern generalizes to other constrained generation tasks.

Repositories

The source code used for the experiments is on GitHub:
`https://github.com/shashuat/anagram-analysis`

References

Burgert, Simon et al. (2023). "Diffusion Illusions: Inverting Text-to-Image Diffusion Models for Visual Illusions". In: *arXiv preprint arXiv:2304.03225*. URL: https://arxiv.org/abs/2304.03225.

Du, Jianfeng et al. (2023). "Learning Energy-Based Compositional Models for Synthesizing Multi-Concept Images". In: *arXiv preprint arXiv:2304.09849*. URL: https://arxiv.org/abs/2304.09849.

Geng, Zheng et al. (2023). "Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models". In: *arXiv preprint arXiv:2311.14421*. URL: https://arxiv.org/abs/2311.14421.

— (2024). "Factorized Diffusion: Generating Compositional Representations of High-Dimensional Data". In: *arXiv preprint arXiv:2403.04877*. URL: https://arxiv.org/abs/2403.04877.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 6840–6851. URL: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Ho, Jonathan and Tim Salimans (2022). "Classifier-Free Diffusion Guidance". In: *arXiv preprint arXiv:2105.05233*. arXiv: 2105.05233 [cs.LG]. URL: https://arxiv.org/abs/2105.05233.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980*. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

Konstantinov, Dmitriy et al. (2023). *DeepFloyd IF: A scalable and high-performance text-to-image model with a cascading architecture*. https://huggingface.co/DeepFloyd/IF-I-L-v1.0. Hugging Face Model Card and Technical Details.

Liu, Saining et al. (2022). "Compositional Conditional Sampling in Diffusion Models". In: *arXiv preprint arXiv:2206.01719*. URL: https://arxiv.org/abs/2206.01719.

Oliva, Aude, Philippe G. Schyns, and Phillipe G. Schyns (2006). "Hybrid Images". In: *ACM Transactions on Graphics (TOG)* 25.3, pp. 527–532. URL: https://web.mit.edu/persci/pub/hybrid-images/oliva_2006_hybrid_images.pdf.

Poole, Ben et al. (2022). "DreamFusion: Text-to-3D using Text-to-Image Diffusion Models". In: *arXiv preprint arXiv:2209.14988*. URL: https://arxiv.org/abs/2209.14988.

Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

Rombach, Robin et al. (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. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Sohl-Dickstein, Jascha et al. (2015). "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. PMLR, pp. 2256–2265. URL: https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Song, Jiaming, Chenlin Meng, and Stefano Ermon (2020). "Denoising Diffusion Implicit Models". In: *arXiv preprint arXiv:2010.02502*. arXiv: 2010.02502 [cs.LG]. URL: https://arxiv.org/abs/2010.02502.

Tancik, Matthew (2023). *Illusion Diffusion*. https://github.com/tancik/IllusionDiffusion. GitHub Repository.

Xu, Yong et al. (2024). "Multi-Task Compositional Sampling for Visual Anagrams". In: *arXiv preprint arXiv:2403.00318*. URL: https://arxiv.org/abs/2403.00318.

Appendix: Image Galleries

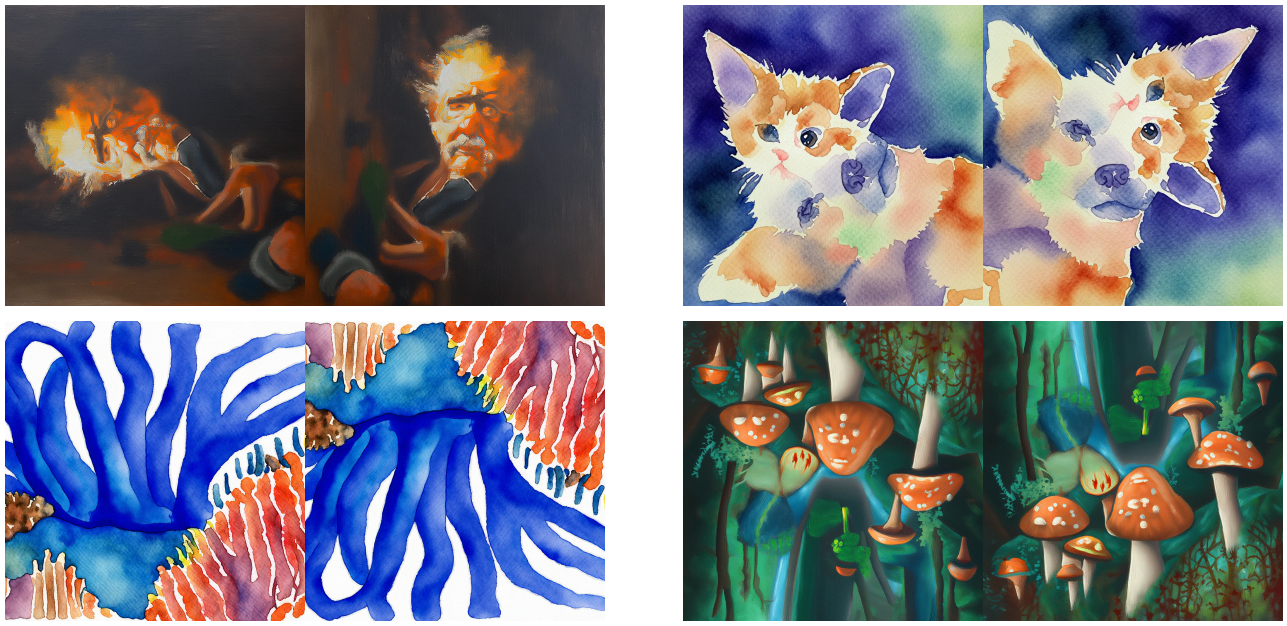


Figure 8. Gallery of Generated Images for Variant 0. The images, displayed in a 2x2 grid (left-to-right, top-to-bottom), correspond to the following prompts: Campfire Man, Kitten Puppy, Reef Octopus, and Wizard Forest.

Figure 9. Gallery of Generated Images for Variant 1. The images, displayed in a 2x2 grid (left-to-right, top-to-bottom), correspond to the following prompts: Campfire Man, Kitten Puppy, Reef Octopus, and Wizard Forest.

Figure 10. Gallery of Generated Images for Variant 2. The images, displayed in a 2x2 grid (left-to-right, top-to-bottom), correspond to the following prompts: Campfire Man, Kitten Puppy, Reef Octopus, and Wizard Forest.

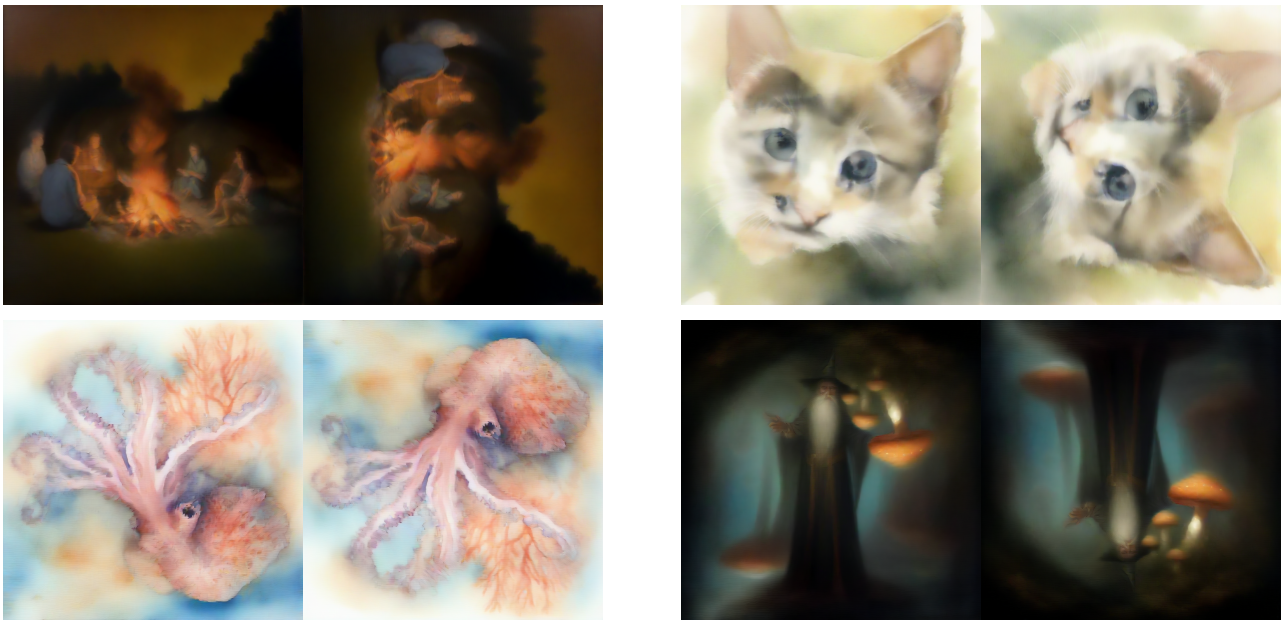


Figure 11. Gallery of Generated Images for Variant 3. The images, displayed in a 2x2 grid (left-to-right, top-to-bottom), correspond to the following prompts: Campfire Man, Kitten Puppy, Reef Octopus, and Wizard Forest.