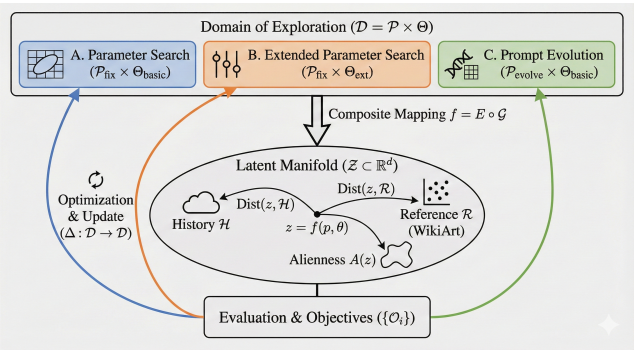
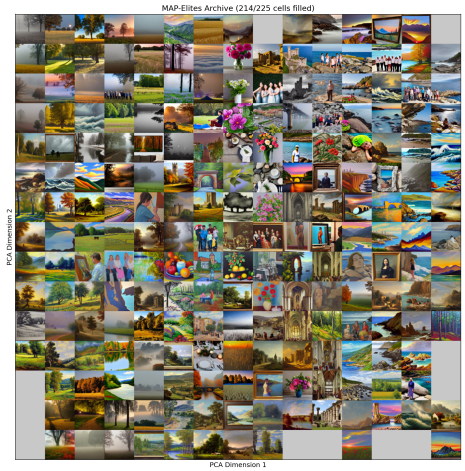
Looking for the Alien Art: On/Off the Manifold

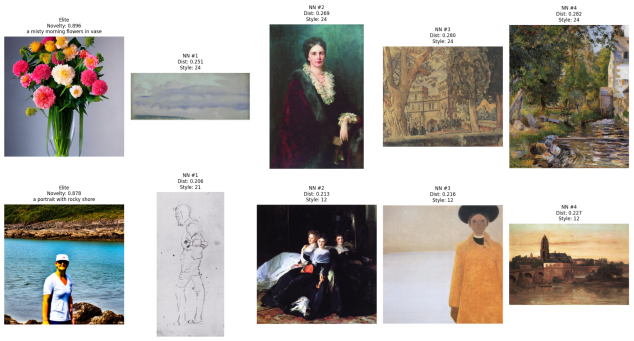
Avi Oberoi
University of Chicago
aoberoi1@uchicago.edu



(a) Exploration Pipeline



(b) MAP-Elites archive in learned feature space



(c) Novel elites vs. WikiArt neighbours

(d) Elite evolution: early vs. late iterations

Figure 1. Qualitative overview of our exploration: (a) Overall Pipeline (Courtesy: Google's Nano Banana Pro); (b) MAP-Elites archive with DINO and art-constrained prompts, each cell shows the elite image for one PCA bin, the archive fills 95.1% cells, indicating broad coverage of the painting manifold; (c) Novel Elites and their nearest WikiArt neighbours, illustrating off-manifold yet interpretable images; (d) Qualitative samples from early (top row) and late (bottom row) iterations show that novelty is maintained while prompts drift from landscapes towards more structurally diverse scenes.

Abstract

We are interested in finding out what drawings, arts, pictures, or images, as a whole, would look like that are alien to us as humans. We began with the latent space of the Foundational Models (FMs) as our exploration ground, with the prior assumption that the representations learned by the FMs approximate the statistical manifold of images shared by humans, and that the models would also find the same things surprising as humans, given their training datasets.

We study this question for Stable Diffusion v1.5 by coupling it with a novelty metric in the DINO and CLIP feature spaces and a geometric reference built from 81k WikiArt images. We first quantify how far naive random sampling and CMA-ES optimization of generation parameters (seed, guidance scale, steps, and internal controls) move the model in this space. Both strategies yield only modest gains in novelty and remain confined to a narrow semantic basin defined by the prompt, despite ∼30-40% improvements over a random baseline. We then apply MAP-Elites in prompt space, using DINO embeddings as the behavior descriptor. This quality-diversity search covers over 95% of a 15×15 archive, doubles mean novelty relative to random search, and places roughly two-thirds of elites in low-density regions of the WikiArt manifold according to multiple geometric criteria. A CLIP-based variant achieves substantially lower coverage and novelty, indicating that language-tied embeddings encourage staying on-manifold.

1. Introduction

What would art from aliens look like? We do not have examples and we do not even know which concepts we are missing. The problem lives in the domain of unknown unknowns. We can use early expeditions as a metaphor to put the problem into perspective.

Early explorers did not know what new continents looked like in advance. They had a boat, some instruments, and a rough model of the Earth. We are in a similar position. Our "ocean" is the model manifold: the embedding space induced by large vision and multimodal models [2, 21]. Our instruments are distances and densities in that space. Our boat is a generative model that lets us move around the manifold by changing prompts and internal controls.

The first question is what we mean by "alien" in this setting. For humans, many different things can feel alien: a rare deep-sea organism, an artifact from an unfamiliar culture, a visual illusion that breaks a familiar rule. Representationally, these are points that our visual system has seen rarely or not at all. We borrow that view for learned models. Given an embedding space trained on human image data, an image is alien if it lies in a low-density or off-manifold region, while still being interpretable as an image [7]. Pure noise can be surprising to a model in the sense of having high loss, but for humans most noise looks the same. We therefore explicitly exclude trivial noise and obviously broken samples from our notion of alien art.

This connects to how we think about representations more broadly. Cognitive science distinguishes prototype and exemplar theories of categories, and information theory gives us ways to talk about surprise. We could follow those threads in depth, but here we make one working assumption: there is some underlying structure of visual reality that can be captured in a representation, whether in language, logic, laws of nature, axioms, or learned embeddings. A recent work suggests that large Foundation Models (FMs) are converging toward such a shared structure: as vision and language models scale, they tend to agree more on which datapoints are similar, and different architectures trained on different datasets produce embeddings with similar geometry [10]. We treat this as a pragmatic, Platonic hypothesis rather than a claim about the mind, i.e., there is at least one statistical manifold of images that these models approximate well enough to use as a common coordinate system.

If that manifold is our ocean, where should we look for alien art? Two obvious places are the tails of the distribution and directions that lie mostly outside the variation seen in human art. In other words: low-density regions and off-manifold directions in FMs embedding space. We also want a human-aligned notion of alienness. If a model finds an image surprising but that surprise is entirely due to a known failure mode, such as adversarial perturbations that humans perceive as random noise, then the result is not useful for us. We care about images that are unlikely under both the generator and a large corpus of human artworks, yet still read as coherent scenes.

Given this view, the question becomes: how should we steer? Our boat is Stable Diffusion [23]. Its parameters include the text prompt, sampler seed, classifier-free guidance scale, step count, and a small set of plug-and-play diffusion knobs that control internal attention and feature strength [8, 27]. We define novelty scores over embeddings from CLIP and DINO [2, 21]: historical novelty with respect to an evolving archive, and reference distance to a fixed set of human artworks (WikiArt) in the same embedding space [1, 26]. These scores tell us whether we are revisiting familiar waters or drifting into low-density regions.

Search on this landscape is a black-box problem. Gradients with respect to high-level controls such as the prompt or plug-and-play knobs are hard to obtain and expensive to trust. At the same time, the space is structured enough that naive random sampling is wasteful. Evolutionary methods sit in this middle ground. Novelty search and quality-diversity algorithms such as MAP-Elites were designed to explore non-differentiable, deceptive spaces by rewarding behavioral diversity rather than a single scalar objective [13, 17]. They are also practical for our setting: they only require function evaluations, they parallelize well, and they fit within our class project budget.

We therefore study three families of steering strategies. First, random search over the parameters with a fixed prompt. Second, covariance-matrix adaptation (CMA-ES) over the parameters: a state-of-the-art evolutionary strategy that steers in directions that increased novelty in the past [6]. Third, quality-diversity search over prompts and parameters using MAP-Elites-style archives. Here language becomes

part of the navigation: prompt mutations move us between semantic basins, and the archive keeps track of diverse high-quality elites across behavior space.

Our experiments ask which combinations of prompts, internal controls, and search algorithms best satisfy this constraint. They also ask a structural question: to what extent can we illuminate low-density and off-manifold regions of the foundation model manifold without changing the models themselves?

2. Related Work

We build on three threads of prior work: representation manifolds in foundation models, methods for controlling diffusion generators, and search algorithms designed for exploration rather than pure optimization.

2.1. Representation manifolds and foundation model embeddings

Large vision and multimodal models such as CLIP and DINO are now widely used as generic feature extractors.[2, 20, 21] Their embeddings support zero-shot classification, retrieval, and text-image alignment, and they correlate with human similarity judgments in many settings.[10, 21] As these models scale, different architectures trained on different datasets tend to agree more on which images are close. Recent work has formalized this as a Platonic representation hypothesis. [10]

This view has roots in vision science. Marr's theory of vision treats perception as a cascade of representations, from the primal sketch through 2.5D and 3D descriptions, each organizing the visual world into structured descriptions rather than raw pixels.[15] Koenderink et.al. describe visual space as a manifold: a continuous medium where local relationships and small deformations matter more than absolute coordinates.[11] Treating CLIP or DINO embeddings as points on a manifold is consistent with this perspective. It gives us a concrete space in which we can talk about neighborhoods, tails, and directions.

Classic work on categorization contrasts prototype theories, where categories are represented by central tendencies, with exemplar theories, where they are represented by stored examples.[18, 24] Both views agree that typical category members cluster densely in representation space, while rare or atypical exemplars live in the tails. In this paper we adopt the same geometric intuition. We treat "alien" images as those that lie far from the dense cores of human art datasets and far from the regions our generator typically visits, while still forming coherent scenes.

Several works use FMs embeddings as scoring functions. CLIP guidance for image generation treats the similarity between an image and a text prompt as a differentiable reward.[3, 21] Kumar et al. introduced Automating the Search for Artificial Life (ASAL) algorithm for measuring diversity and novelty in ALife simulation datasets, and uses FMs embedding scores to drive open-ended exploration.[12] We are directly inspired by this line of work, and our contribution is to adapt the same idea to the space of generated artworks: we use embeddings to measure distance from our own search history and from a large reference corpus of human art, and we study which search strategies best populate low-density regions under these metrics.[1]

2.2. Controlling diffusion models

Diffusion models support many forms of control. Adding and removing noise along the sampling trajectory enables image editing, inpainting, and conditional generation without full retraining.[14, 16] Conditioning mechanisms range from concatenating extra channels to using cross-attention over text tokens or spatial feature maps.[23] Plug-and-play diffusion features expose intermediate activations and attention maps as control points for structure and layout, showing that internal features can be used as knobs for text-driven image-to-image translation.[27]

On the parameter side, fine-tuning techniques such as low-rank adaptation (LoRA) and ControlNet add new concepts, styles, or conditioning signals on top of a frozen backbone.[9, 28] These methods enlarge the space of images a given diffusion model can realize and give users more levers for creative control. In our work, we treat a small set of existing sampler and plug-and-play knobs as a control vector and ask how different search strategies populate the embedding manifold.

2.3. Generative models for creative exploration

Diffusion models have quickly been adopted for creative exploration. Systems such as ConceptLab use vision-language models both as generators and as instruments, probing how concepts combine and where surprising images appear in representation space.[1, 3] Other recent work uses diffusion pipelines and foundation model scores to search for unusual or aesthetically pleasing images, sometimes with human feedback in the loop or with constraints on style and prompt adherence. A parallel thread uses quality-diversity algorithms and foundation model assessors to evolve diverse images or text-image pairs across a behavioral space[17].

Our focus is narrower and more geometric. We restrict ourselves to a single backbone generator and to automated search over prompts and control parameters. Within that setting we ask which combinations of embeddings and search algorithms are best suited to uncover low-density or off-manifold images, and how far we can go before generation quality breaks down.

2.4. Evolutionary and quality-diversity search

Evolutionary algorithms provide a natural toolkit for black-box exploration. Covariance Matrix Adaptation (CMA-ES)

adapts multivariate normal search distribution to maximize a scalar objective in continuous spaces and is widely used as a strong baseline for derivative-free optimization.[6] Novelty search and quality-diversity algorithms such as MAP-Elites modify this picture. Instead of pushing toward a single optimum, they reward behavioral diversity and aim to illuminate a space of possible solutions.[13, 17] The search maintains an archive of elites, one per region of a behavior descriptor space, and continually fills gaps.

These methods have been applied in robotics, games, and, more recently, in creative domains. Prior work has used MAP-Elites and related algorithms to evolve diverse images, prompts, or behaviors, often using CLIP-based scores as fitness or novelty signals.[12, 17] Our setup is closely aligned with this line. We use CMA-ES as a baseline for scalar "alienness" objectives such as combined novelty and reference distance. We use MAP-Elites-style archives to encourage coverage of the embedding manifold and to sustain exploration rather than converging to a single high-scoring mode.

2.5. Gradient-based search and adversarial artifacts

An alternative to evolutionary search is to differentiate through the generator and optimize inputs by gradient descent. CLIP-guided generation, feature visualization, and adversarial attacks on vision models all follow this pattern: define a loss on the embedding of the generated image, backpropagate, and update pixels or latents.[3, 5, 19] These methods can produce highly optimized images that strongly activate model features, but they are also prone to adversarial artifacts. Images that look like pure noise or uninterpretable texture to humans can achieve extreme scores.

2.6. Cognitive views on surprise and alienness

Information theory often defines surprise as the negative log probability of an observation under a predictive model.[25] Predictive processing accounts of perception likewise treat the brain as minimizing prediction error, with unexpected events that violate learned regularities playing a special role.[4, 22] In this language, alien images are those that have low probability under both the model's internal generative distribution and under the empirical distribution of human art.

Our definition of alienness always includes a coherence constraint: we only count images that remain interpretable as scenes or objects. This constraint shapes both our metrics and our choice of search algorithms in the remainder of the project.

3. Methods

In our experimental setup, we fix a generative backbone, define controllable parameters θ, specify the embedding spaces and scores we use to approximate novelty and off-manifoldness, and describe the search procedures that act on θ.

3.1. Generative backbone and control parameters

All images are generated with a Stable Diffusion v1.5 text-to-image model via the diffusers library. We use 512×512 resolution, a DDIM sampler, disabled safety checker, and attention slicing on a single GPU.

We expose a low-dimensional control vector

$$\theta = (\text{seed}, \text{cfg\_scale}, \text{steps})$$

and, in some experiments, an extended vector

$$\theta_{\text{ext}} = (\text{seed}, \text{cfg\_base}, \text{steps}, \text{cfg\_slope}, \text{eta\_range}, \text{prompt\_strength})$$

The basic controls are:

- **Seed**: $[0, 10^6]$ that initializes the diffusion RNG.
- **Classifier-free guidance scale** cfg_scale: $[3.0, 15.0]$ controlling prompt adherence.
- **Steps**: number of inference steps $\{20, 25, 30, 40\}$.

The extended controls add simple plug-and-play style perturbations:

- **DDIM Stochasticity** $\eta \in [0.0, 1.0]$: Controls the stochasticity of the sampling process, interpolating between deterministic DDIM ($\eta = 0$) and DDPM-like behavior ($\eta = 1$).
- **Guidance Rescale** $\phi \in [0.0, 0.7]$: Rescales the noise prediction to mitigate over-exposure artifacts caused by high CFG scales.
- **CLIP Skip** $k \in \{0, 1, 2\}$: Uses hidden states from the $(L - k)$-th layer of the CLIP text encoder, yielding less abstract semantic interpretations.
- **Prompt Strength** $\alpha \in [0.7, 1.3]$: Linearly scales the text embedding vector, modulating the strength of the conditioning signal relative to the unconditional embedding.

Parameters are normalized to $[0, 1]^d$ inside the optimizers and denormalized before sampling.

3.2. Embedding spaces and scoring

We work in two embedding spaces:

- **CLIP**: ViT-L/14 encoder, 768-dimensional image embedding, L2-normalized. This space is aligned with text prompts and is used mainly for behavior descriptors and ablations. [21]
- **DINOv2**: DINOv2-base ViT encoder, 768-dimensional CLS token, L2-normalized. This space emphasizes visual structure and is our primary space for novelty and distance-to-art measurements. [20]

3.2.1. Historical novelty

Let $z \in \mathbb{R}^d$ be the embedding of a generated image and $H = \{h_1, \ldots, h_N\}$ the set of embeddings produced so far. We define a k-NN novelty score

$$\text{novelty}(z; H) = \frac{1}{k} \cdot \sum_{i=1}^{k} \big(1 - \cos(z, h_{(i)})\big),$$

where $h_{(1)}, \ldots, h_{(k)}$ are the k nearest neighbors of z in H under cosine similarity. We use $k = 10$ and update H online. FAISS is used when H grows beyond a few thousand points.

3.2.2. Distance to human art

To measure distance from human art, we build a reference set $R = \{r_1, \ldots, r_M\}$ of WikiArt paintings [26] embedded with the same encoder. For a generated embedding z we define

$$\text{ref\_dist}(z; R) = \frac{1}{k} \cdot \sum_{i=1}^{k} \big(1 - \cos(z, r_{(i)})\big),$$

where $r_{(i)}$ are the k nearest neighbors in R. High values indicate that z lies in a sparse region relative to typical art styles and subjects.

3.2.3. Combined alienness

For scalar optimization we sometimes mix the two scores into a single "alienness" score

$$A(z) = \alpha \cdot \text{novelty}(z; H) + (1 - \alpha) \cdot \text{ref\_dist}(z; R),$$

with $\alpha = 0.5$ unless stated otherwise. Later we complement this scalar measure with global geometric diagnostics.

3.3. Search procedures

We compare four families of search procedures over θ and, in one case, over prompts. All share the same generator and encoders.

3.3.1. Random search baseline

Random search samples θ uniformly within the allowed ranges for seed, guidance scale, and steps. For each θ we generate images, embed them with DINO (or CLIP in ablations), compute novelty against H, and append z to H. This gives an empirical novelty distribution for naive exploration under a fixed prompt.

3.3.2. CMA-ES on basic parameters

We then replace uniform sampling by Covariance Matrix Adaptation Evolution Strategy (CMA-ES) over θ. We wrap θ in a normalized cube $[0, 1]^3$, initialize a Normal search distribution with mean at the center and step size $\sigma_0 = 0.3$, and use $\lambda \approx 12$ samples per iteration.

At each CMA-ES step we sample a batch of θ values, generate images, evaluate novelty or $A(z)$, and feed the scores back to CMA-ES to update its mean and covariance. This tests how far a strong optimizer can push novelty when it only controls the standard sampler knobs.

3.3.3. CMA-ES with extended controls

To check whether richer plug-and-play controls change this picture, we repeat the CMA-ES setup over θ_{ext}. The search now operates in seven dimensions, adding DDIM stochasticity (η), guidance rescaling, CLIP layer skipping, and prompt strength scaling. Stochasticity and guidance rescaling are applied during the sampling loop, while CLIP skipping and prompt strength modify the text conditioning embeddings prior to generation.

The objective and update loop are unchanged.

3.3.4. MAP-Elites with prompt evolution

Finally, we expand the search space to include language. Each candidate is a pair (p, θ), where p is a text prompt and θ is a small set of sampler controls. We run a MAP-Elites-style algorithm that populates a 2D behavior space with diverse elites.

Behavior spaces.. We use two parameterizations:

- A semantic-axis mode: define axes in CLIP space by embedding anchor phrases (for example, "abstract art" vs. "figurative painting"), project each image embedding onto these directions, and discretize into a grid of cells.
- An automatic PCA mode: run PCA on accumulated DINO embeddings and use the first two principal components as coordinates, updating the PCA fit periodically as the archive grows.

Each cell in the grid stores at most one elite.

Art-constrained prompt evolution.. Seed prompts describe classic painting subjects such as landscapes and seascapes. Mutation operators then edit:

- Subject matter (landscapes, portraits, still lifes, architectural scenes),
- Adjectives and mood (for example, melancholic, luminous, mysterious),
- Style and medium (movements and artist references; oil, watercolor, pastel),
- Time, lighting, and weather (dawn, golden hour, stormy skies).

Additional operators introduce controlled conceptual tension:

- Physics and environment twists (non-Euclidean space, sideways gravity, deep space),
- Concept fusions and scientific twists (for example, deep sea biology with gothic architecture),

- Mundane anchors (office cubicle, vending machine, dentist waiting room).

All mutations keep prompts within an art-compatible schema so that outputs remain interpretable paintings rather than trivial noise.

MAP-Elites loop.. The archive maps behavior-space cells to elites $e = (p, \theta, z, s)$, where z is the embedding and s is a scalar score (novelty, ref_dist, or $A(z)$). Each iteration:

1. Samples a parent elite (or a seed prompt early on).
2. Mutates its prompt and optionally perturbs θ.
3. Generates an image, computes z, its behavior coordinates, and scores.
4. Inserts the candidate into its cell if the cell is empty or if the new score is higher.

We track archive coverage and the score distributions to see how well MAP-Elites spreads out over the embedding space and into low-density regions.

3.4. Off-manifold geometry

The scores above are local. To study global geometry relative to human art, we analyze distances in WikiArt space.

3.4.1. Mahalanobis distance

Let μ be the mean in this space and Σ a shrinkage covariance estimate. For a generated embedding z we compute its Mahalanobis distance

$$D_M(z) = \sqrt{(z-\mu)^\top \Sigma^{-1} (z-\mu)}.$$

We then compute the percentile of $D_M(z)$ with respect to the WikiArt distance distribution. High percentiles indicate directions where human art is rare.

3.4.2. Low-variance subspace activity

To focus on directions that WikiArt barely uses, we take the bottom m principal components ($m = 10$) and measure

$$\text{low\_pc\_norm}(z) = \left\| V_{\text{low}}^\top (z-\mu) \right\|_2.$$

Again we convert this to a percentile relative to WikiArt. Large values flag embeddings that put significant mass into low-variance directions.

3.4.3. Alien quadrants and set-level scores

For visualization we define an "alien quadrant" in the plane spanned by novelty and ref_dist: images whose novelty and ref_dist both exceed the median over the set. We intersect this quadrant with thresholds on Mahalanobis percentile and low_pc_norm percentile to obtain a subset of images that are novel to our search, far from typical art, and geometrically atypical.

In the next section we instantiate these procedures with concrete runs and report how each behaves across embedding spaces and metrics. The code for the project can be found at https://github.com/avioberoi/alien-art-search

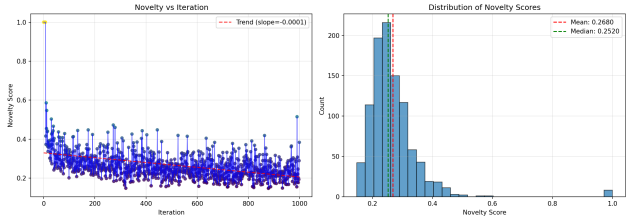


Figure 2. **Baseline random search with a fixed prompt.** Stable Diffusion is run for N=1000 samples with prompt *"a painting of a landscape"* and random $\theta = \{\text{seed}, \text{cfg}, \text{steps}\}$. Left: novelty score vs. iteration; right: histogram of novelty scores. After the initial spike (history is empty), novelty collapses to a tight band around $\mu \approx 0.27$ and shows a slight downward trend, indicating that random search explores only a small region of the embedding space.

4. Experiments and Results

Unless otherwise noted, all runs use SD v1.5 at 512×512 resolution, DINOv2 image embeddings, and the k-NN novelty and WikiArt reference distances defined in Section 3.

4.1. Baseline: Random Walk in Parameter Space

We first fix a mundane prompt, *"a painting of a landscape"*, and sample only the generation parameters

$$\theta = \{\text{seed},\ \text{cfg},\ \text{steps}\} \tag{1}$$

For this setting we generate N=1000 images with θ drawn uniformly from the ranges in Section 3.

Using DINO as the feature space and our k-NN novelty metric, the resulting novelty distribution is

$$\mu = 0.268, \quad \text{median} = 0.252, \quad \sigma = 0.0905$$

Figure 2 shows novelty as a function of iteration together with its histogram.

Figure 3 shows the novel images from this run. All are recognizably oil landscapes: hills, trees, rivers, and skies rendered with different palettes, frames, and brushwork. Some compositions are slightly exaggerated or stylized, but nothing in this gallery would be judged "alien" by a human viewer. Varying only θ changes *how* the painting is rendered, not *what* is being depicted.

To understand whether this collapse is a limitation of the search procedure or of the novelty lens, we repeat random sampling for N=100 images with a more adventurous prompt, *"abstract alien landscape, surreal, otherworldly"*, and evaluate novelty once with CLIP and once with DINO on the *same* images. Table 1 summarizes the statistics.

Despite identical inputs, CLIP sees most of these samples as minor variations of a single concept, "alien landscape," and assigns low novelty. DINO, trained without text supervision, is more sensitive to visual and structural differences.

Figure 3. **Most novel samples under baseline random search.** Top images ranked by DINO-based novelty from the run in Figure 2. All images remain within the space of conventional landscape paintings, with variation in framing, color, and brushwork but no clear category violations. Randomly walking θ explores stylistic degrees of freedom inside a single semantic basin rather than uncovering truly alien structure.

Encoder	Min	Max	Mean	Median	Std
CLIP	0.035	1.000	0.108	0.087	0.130
DINO	0.198	1.000	0.370	0.362	0.121

Table 1. **Random search with an "alien" prompt under different embedding spaces.**

Table 2. **CMA-ES over generation parameters with fixed prompt.**

Method	Param. space	Mean ↑	Std.
Random search	θ	0.268	0.091
CMA-ES (basic)	θ	0.364	0.106
CMA-ES (expanded)	θ_{ext}	0.385	0.107

This divergence motivates the rest of the paper: when the goal is to detect visually alien structure rather than semantic category shifts, DINO provides a more informative manifold than CLIP.

4.2. Optimizing Parameters Without Changing the Prompt

We optimize the same prompt, *"a painting of a landscape"*, using CMA-ES over (i) the basic parameter set $\theta = \{\text{seed}, \text{cfg}, \text{steps}\}$ and (ii) the expanded set θ_{ext} that also includes internal diffusion controls (DDIM stochasticity, Guidance Rescale, and prompt strength).

Figure 4 summarizes the runs. Both variants quickly find higher-scoring regions than random search, but the search remains confined to the same semantic basin: landscapes stay landscapes. The expanded control space yields only a modest gain in novelty.

The galleries in Figures 5 and 6 show the three most novel images under each setting. They are high-quality oil landscapes with more aggressive lighting, framing, and color choices, but they remain well within the familiar "painting of a landscape" category. CMA-ES is an effective optimizer on this surface; it does not, by itself, invent alien art.

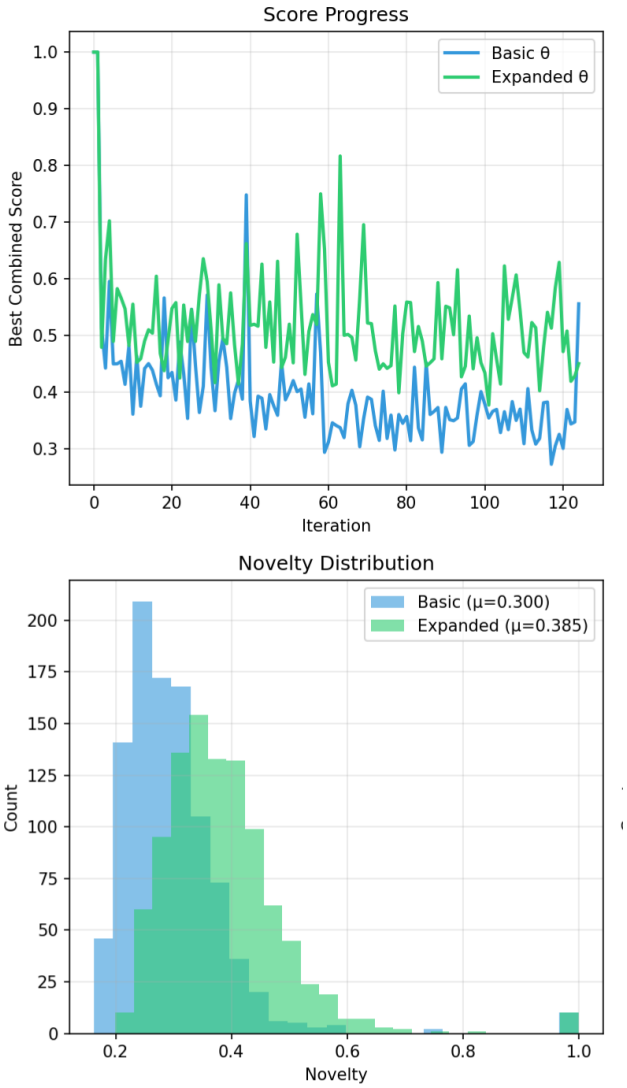


Figure 4. **CMA-ES on basic vs. expanded parameter sets for a fixed prompt.** Top: best combined score over iterations. Bottom: novelty distributions. Both variants improve over random search, and the expanded controls yield a small additional gain, but the overall novelty band remains narrow.

4.3. Prompt-space quality diversity with MAP-Elites

We now let the search operate in prompt space. MAP-Elites maintains an archive over a 2D behavior grid defined by the first two principal components of DINO embeddings of all elites (Section 3). Each cell stores the highest-scoring image seen at that location.

Art-constrained DINO run. We first restrict prompts to an "art-like" schema (paintings, canonical subjects, art movements). Figure 1(b) shows the final archive: MAP-Elites fills

Figure 5. **Top-3 most novel images for CMA-ES on the basic parameter set θ (fixed prompt).** All samples are recognizable oil landscapes with more extreme compositions, yet none would be judged "alien" by a human viewer.

Figure 6. **Top-3 most novel images for CMA-ES on the expanded parameter set θ_{ext}.** Internal diffusion controls intensify color, contrast, and structure but do not move the images outside the landscape manifold.

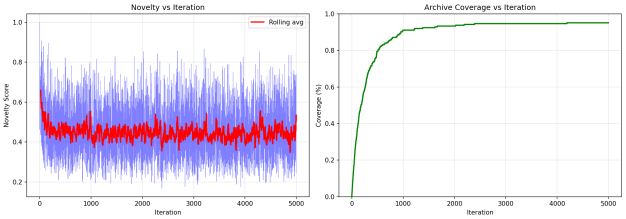


Figure 7. **DINO + art-constrained MAP-Elites.** Left: novelty vs. iteration (blue) with a rolling average (red). Right: archive coverage vs. iteration. Novelty stays in the 0.4-0.5 band while coverage quickly saturates near 95%.

214/225 cells (95.1%), covering the behavior space with diverse yet recognizably painterly images. Novelty remains high over 5,000 iterations, and archive coverage saturates early (Fig. 7). The "alien vs. typical" comparison in Fig. 8 illustrates the effect of the novelty score: high-novelty elites are often category-violating paintings

that sit far from the WikiArt manifold, while low-novelty examples collapse to conventional landscapes.

Dropping the art constraint. Next we allow the full prompt-mutation bank, including mundane anchors (e.g., "grocery store", "vending machine"). The archive in Fig. 9 reaches 207/225 filled cells (92.0%). The top-novel elites (Fig. 10) now include hybrids of landscapes with supermarket facades, vending machines, filing cabinets, and other banal objects. This shows that once the prompt prior is relaxed, MAP-Elites eagerly occupies low-density regions by combining the painting manifold with everyday photographic motifs. But this might be true because of choice of prompts.

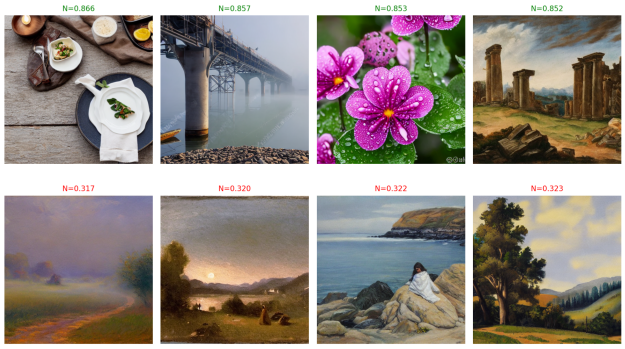


Figure 8. **Alien vs. typical elites under DINO.** Top row: high-novelty images. Bottom row: low-novelty neighbors from the archive. The metric systematically prefers visually unusual, category-violating structure over conventional landscapes.

Figure 9. **DINO + unconstrained prompts.** MAP-Elites still attains high coverage (207/225 cells), but many niches are populated by hybrids of classical painting structure with mundane objects (e.g., vending machines, storefronts, household appliances).

CLIP vs. DINO as the search manifold.. Finally we repeat MAP-Elites with CLIP ViT-L/14 as the embedding space while keeping the same prompt evolution. The CLIP archive in Fig. 11 fills only 86/225 cells (38.2%), and the novelty trace collapses into a narrow band around conventional semantics (Fig. 12). Elites are dominated by photorealistic portraits, weddings, and typical social scenes (Fig. 13), rather than the structurally odd images seen with DINO. Table 3 summarizes the two main settings. CLIP behaves like a strong semantic prior, keeping the search on-manifold, while DINO exposes off-manifold visual structure and supports

Figure 10. **Top novel elites for the unconstrained DINO run.** The highest-scoring images often blend painterly composition with everyday scenes: grocery stores, storage closets, clocks, office furniture. Novelty is achieved by colliding the art manifold with banal categories.

Figure 11. **MAP-Elites archive with CLIP embeddings.** Only 86/225 cells (38.2%) are filled. The archive concentrates around a semantic core of people, portraits, and everyday scenes, indicating that CLIP acts as a strong semantic anchor.

Embedding	Coverage	Mean novelty	Off-manifold
DINO	95.1%	0.574	63.7%
CLIP	38.2%	0.297	42.9%

Table 3. **Summary of MAP-Elites runs.** DINO yields higher coverage, higher novelty, and a larger fraction of elites in low-density regions of the WikiArt manifold than CLIP.

much richer quality diversity.

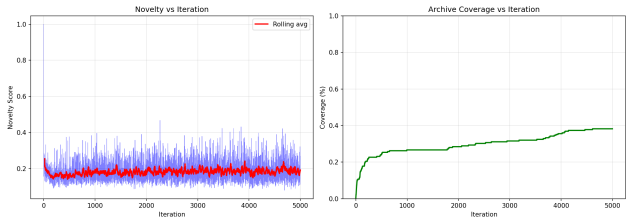


Figure 12. **CLIP-based MAP-Elites.** Left: novelty vs. iteration, which quickly settles into a tight band. Right: archive coverage vs. iteration, plateauing around 38%. Compared to Fig. 7, exploration is much more limited.

Figure 13. **Top novel elites under CLIP.** High-scoring images remain semantically conventional (portraits, weddings, group photos), despite prompt evolution. CLIP rewards "expected surprise" within known categories rather than the category-violating forms favored by DINO.

4.4. Illuminating Low-Density Regions of the Art Manifold

We now ask whether the MAP-Elites archive lives on the same "art manifold" as WikiArt, or whether a substantial fraction of elites occupy low-density, off-manifold regions.

Archive geometry. The DINO-based MAP-Elites run produces 214 elites on a 15×15 grid (95.1% coverage). Figure 14 shows when each cell is first filled and the final novelty per cell, together with archive coverage over time. Cells near the center of the PCA grid are filled early and repeatedly refined; coverage approaches 90% within the first 10^3 iterations and slowly climbs towards saturation thereafter. Early and late elites maintain comparable novelty (correlation -0.16 with iteration; early mean 0.60, late mean 0.55), as illustrated in the qualitative trajectories of Fig. 1(d).

Low-density illumination. To relate the archive to human art, we embed 81,444 WikiArt paintings in the same DINO space and measure each elite's average k-NN distance to WikiArt. Figure 15 summarizes these statistics. Elites have mean distance 0.37 with median percentile 82.8; 63.7% lie beyond the 75th percentile of WikiArt–WikiArt distances and 41.2% beyond the 90th percentile. Nearest-neighbor galleries (not shown) confirm that high-distance elites correspond to images that are visually plausible yet stylistically

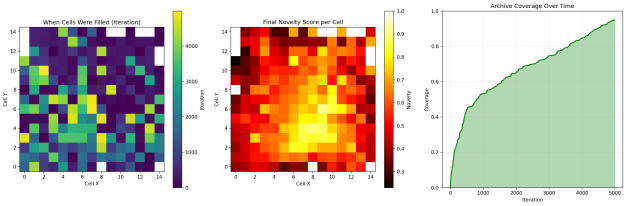


Figure 14. **Cell dynamics for the DINO MAP-Elites archive.** Left: iteration at which each cell is first filled. Middle: final novelty score per cell. Right: archive coverage over time.

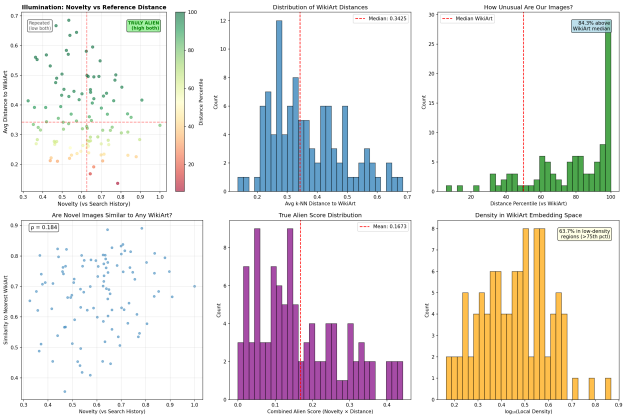


Figure 15. **Illumination of low-density regions.** DINO-based novelty (vs. search history) versus distance to WikiArt, the distribution of distances and percentiles, and the resulting "true alien" score reveal that a majority of elites live in low-density regions of the WikiArt manifold.

Metric	Mean ± SD	% > 75th pct.
Mahalanobis distance	33.1 ± 7.4	61.8
k-NN distance (DINO)	0.37 ± 0.12	63.7
Combined alien score	0.66 ± 0.17	—

Table 4. **Summary of off-manifold statistics for MAP-Elites elites relative to WikiArt.**

or structurally unusual compared to their closest WikiArt counterparts.

Geometric off-manifold analysis. To probe geometry beyond local density, we fit a PCA+Ledoit-Wolf model to WikiArt and compute, for each elite, Mahalanobis distance, PCA reconstruction error, and activity in low-variance directions. Figure 16 shows that elites are systematically farther from the WikiArt distribution than typical paintings along all three axes; over 60% exceed the 75th percentile of WikiArt Mahalanobis distances. Novelty has only weak correlation with any single geometric metric (e.g., $\rho = -0.095$ for Mahalanobis distance), indicating that our search discovers images that are both historically novel (vs. the archive) and statistically off-manifold (vs. WikiArt), but that these two

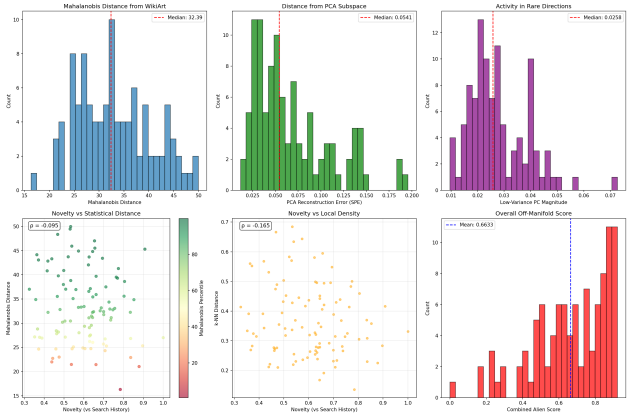


Figure 16. **Off-manifold geometry of MAP-Elites elites.** Histograms and scatter plots of Mahalanobis distance, PCA reconstruction error, low-variance PC activity, and local density show that a large fraction of elites lie far from the fitted WikiArt manifold.

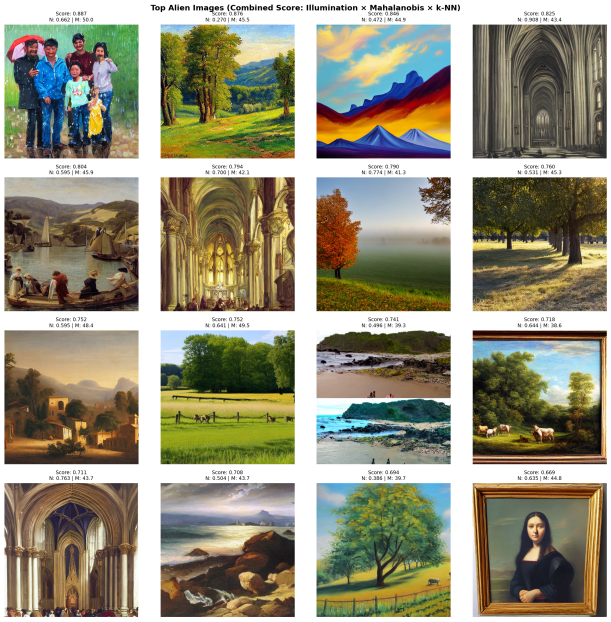


Figure 17. **Top "alien" elites.** Representative images with high novelty and high statistical distance from WikiArt illustrate the kinds of category-violating structures discovered by the system.

notions of "alien" are partially complementary.

References

[1] Panos Achlioptas, Maks Ovsjanikov, K Haydarov, et al. ArtEmis: Affective language for visual art. In *Advances in Neural Information Processing Systems*, 2021. 2, 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2, 3

[3] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders. *Neural Computing and Applications*, 2022. Originally arXiv:2106.14843. 3, 4

[4] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. 4

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 4

[6] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. 2, 4

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*, 2017. 2

[8] Amir Hertz, Ron Mokady, Kfir Aberman, Jay Tenenbaum, Yael Pritch, Ohad Fried, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models, 2021. 3

[10] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20617–20642. PMLR, 2024. 2, 3

[11] Jan J. Koenderink, Andrea J. van Doorn, and Johan Wagemans. Part and whole in pictorial relief. *i-Perception*, 6(6), 2015. 3

[12] Akarsh Kumar, Chris Lu, Louis Kirsch, Yujin Tang, Kenneth O. Stanley, Phillip Isola, and David Ha. Automating the search for artificial life with foundation models, 2024. arXiv preprint. 3, 4

[13] Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011. 2, 4

[14] Andreas Lugmayr, Martin Danelljan, Alejandro Romero, Fisher Yu, Luc Van Gool, and Radu Timofte. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[15] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982. 3

[16] Chenlin Meng, Robin Rombach, Jingyu Gao, , et al. SDEdit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations*, 2022. arXiv:2108.01073. 3

[17] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 75–82, 2015. 2, 3, 4

[18] Robert M. Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57, 1986. 3

[19] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 4

[20] Maxime Oquab, Théo Darcet, Tete Xiao Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. Preprint. 3, 4

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. 2, 3, 4

[22] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1): 79–87, 1999. 4

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[24] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104 (3):192–233, 1975. 3

[25] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948. 4

[26] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. 2, 5

[27] Narek Tumanyan, Mario Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[28] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3