Looking for the Alien Art:
On and Off the Manifold

Avi Oberoi
University of Chicago
aoberoi1@uchicago.edu

Abstract

Akin to the field of artificial life*, we are interested in finding what the drawings, arts, pictures, or images as a whole would look like that are alien to us as humans. We would start with the latent space of the Foundational Models (FMs) as our exploration ground, with a prior that the representations learned by the FMs are aligned with us, and that they would also find the same things surprising as a human would, because they are trained on our data [3]. We will approach the problem in a self-supervised manner with developed methods in (i) Reinforcement Learning [1, 5], (ii) Derivative-free Optimization [4, 7], and (iii) Visual Reasoning [2]. The measures for defining novelty, surprise, and other concepts would be borrowed from existing literature. We will avoid fine-tuning/training costs while directly optimizing novelty/diversity in DINO/CLIP spaces.*

1. Problem

What are we trying to do? The goal is to find art that might be surprising to us (humans) collectively. Obviously, this can be generalized to any idea space, but for the project, we will focus on the image space.

How is it done today? The methods for exploring novel/surprising states/spaces sit at the heart of Reinforcement Learning, where researchers have tried both intrinsic and extrinsic rewards to guide the exploration. Another developed line of research explores derivative-free evolution strategy for modeling search distribution over parameters. A somewhat underexplored area is using reasoning Language models to guide the exploration (Need to double-check)

What is new in our approach and it will be successful? The framing of our research question in our target space/manifold of images is unique, and given the history of developed methods we would use for our problem, it gives us the confidence that it would be successful.

Who cares? The immediate audience for our work would be the people in the art world. Our success would also attract researchers from varied fields to develop upon the methodology.

What are the risks? The risks revolve around the convergence of our methods, not finding a closed-form solution, or simply too alienating art.

Practicalities? No model training/fine-tuning for Derivative-free Optimization and Visual Reasoning strands of our project. Cost is primarily driven by Stable Diffusion (SD) sampling and FM embedding, which run on a single GPU with batched candidates per generation. Reinforcement learning would require significant training efforts and would likely be expensive if we were to pursue learning an optimal policy. The overall timeline is estimated to be around 2-3 months of dedicated work.

2. Methodology

Keeping the scope of the project limited for the class (TTIC 31270), we would love to experiment with the **Derivative-free Optimization** first. Drawing inspiration from Kumar et al.'s Automating the Search for Artificial Life (ASAL) approach [4], we will use FM representations (DINO/CLIP) as a proxy for human judgments to automatically discover image regions that are both coherent and historically novel. We port ASAL's **open-endedness** (persistent novelty over time) and **illumination** (diverse set coverage) from ALife substrates to controllable diffusion generation. We will adopt Tumanyan et al.'s Plug-and-Play (PnP) Diffusion Features to steer structure/layout [6].

Summary of what we will implement first:

1. Sample a population of θ (PnP knobs: layers, attention scaling, CFG schedule, seeds).
2. For each θ: render images (and/or short sequences), extract FM embeddings.
3. **Objective:** compute illumination (set diversity) or open-endedness (historical novelty) score.
4. Update θ via **Sep-CMA-ES**; for illumination, retain a diverse survivor set.
5. Iterate until convergence/budget; log galleries and metric curves.

2.1. Controllable Diffusion Generation

Inject intermediate **decoder spatial features** and **self-attention maps** from a guidance trajectory to steer structure/layout, controlled by two "injection horizons" f (features) and A (self-attention). We will keep the U-Net weights frozen; control happens by swapping features/attention during sampling (DDIM inversion + guided denoising).

θ **(searchable knobs):** layer set for feature injection, layer set for attention injection, attention scaling, negative-prompt text and weight, CFG schedule, steps/seed, optional per-layer gates. (Tumanyan et al. shows that intermediate features encode localized semantics; self-attention preserves fine layout useful inductive bias for controllable novelty [6].)

2.2. Representation & distances

Extract $z = FM(I)$ using DINO (or CLIP) and measure cosine similarity.

2.3. Objectives

Illumination (diversity set): find a set of θ's whose images are far from their nearest neighbor in FM space (nearest-neighbor diversity).

$$\min_{\theta_i} \mathbb{E}\Big[\max_{\theta_j \neq \theta_i} \langle z(\theta_i), z(\theta_j) \rangle\Big]$$

Open-endedness (temporal novelty): for a trajectory (e.g., denoising snapshots or short multi-frame renders per θ), minimize alignment to the historically closest prior outcome (encouraging persistent novelty over time).

2.4. Search algorithm

Sep-CMA-ES [7] over θ for both objectives (black-box, low effective dimension, robust to non-smoothness). For illumination, we will run a **diversity-preserving** survivor selection (keep a batch with low pairwise alignment). For open-endedness, we will maintain a ring buffer of recent embeddings to compute historical nearest neighbors efficiently.

2.5. Evaluation

Benchmarks:(i) Illumination-coverage on a k-center metric in FM space; (ii) Open-endedness-sustained decrease of nearest-historical alignment over steps; (iii) Ablations on θ (e.g., no features / no attention) mirroring PnP's findings that both are critical for structure-faithful control.

Artifacts: galleries of "discovered" clusters; quantitative plots; write-up on failure modes (e.g., texture-only drift; countered via negative prompts and attention-only phases).

References

[1] Chenjia Bai, Lingxiao Wang, Lei Han, Animesh Garg, Jianye Hao, Peng Liu, and Zhaoran Wang. Dynamic bottleneck for robust self-supervised exploration. In *Advances in Neural Information Processing Systems*, pages 17007–17020. Curran Associates, Inc., 2021. 1

[2] Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging, 2025. 1

[3] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. 1

[4] 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, 2025. 1

[5] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020. 1

[6] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022. 1, 2

[7] Kento Uchida, Teppei Yamaguchi, and Shinichi Shirakawa. Covariance matrix adaptation evolution strategy for low effective dimensionality, 2024. 1, 2