Seeking the Unfamiliar but Memorable: Conceptual Creativity as Meta-Learning

Mengye Ren
Agentic Learning AI Lab, New York University
mengye@nyu.edu

Abstract

What does it mean to create a new concept, rather than retrieve a familiar one? Repeatedly sampling a generative model at the same prompt produces variations with similar styles and typical content. We propose that creativity is the production of stimuli that are unfamiliar to an adaptive observer at first sight, but quickly learnable from a few exposures. We formalize this as a *Creator-Appraiser* pair: a Creator generates a candidate, an Appraiser adapts to it for a few inner-loop learning steps, and the Appraiser's improvement becomes the reward the Creator optimizes through. We instantiate the framework with diffusion as the Creator, an autoencoder Appraiser on MNIST, and a CLIP Appraiser with a low-rank adapter for natural images. The diffusion model remains frozen with no additional language conditioning; the meta-learning gradient is enough to produce both stylistic variations and concept compositions that the base model does not generate on its own.

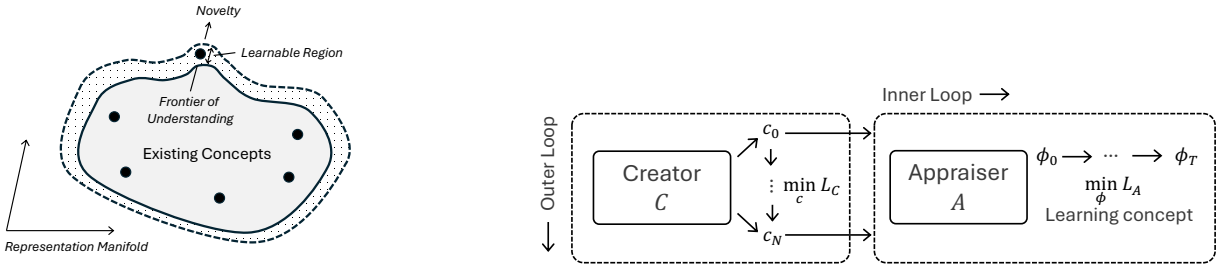
1 Introduction

Humans recognize new concepts from just one or a few examples. Walking into a museum and encountering a Monet painting is enough to begin grasping the "concept" of Impressionism and to recognize its style in other works. This capacity for "few-shot" concept learning is a cornerstone of intelligent behavior (Fei-Fei et al., 2006; Lake et al., 2011), and algorithmic counterparts have advanced rapidly through meta-learning (Finn et al., 2017), prototypical representations (Snell et al., 2017), and in-context learning (Brown et al., 2020).

But this raises a more fundamental question: *what does it mean to create a new concept in the first place?* While humans and machines can readily learn concepts like Impressionism, the creative act itself remains elusive. Situated from the perspective of the creator, how can we map the process of conceptual creation into a computational framework?

Before defining creation, it helps to see that not all novelty is meaningful. An infinite stream of white-noise images is mathematically distinct sample by sample, yet each is perceptually indistinguishable from the next: they are new but offer no structure to grasp. For a new concept to be valuable, it must be grounded in our existing representations of the world; it must, on some level, be understandable from past experience.

Conversely, some things are perfectly understood but uninteresting: a line, a square, or the sequence $1, 2, 3, \ldots$ are entirely comprehensible, yet lack the novelty that makes a creative act. A new concept worth creating must therefore sit between these extremes—grounded enough to be assimilable, yet carrying enough initial perplexity to engage the observer.



(a) Frontier of understanding

(b) Creator-Appraiser Meta-Learning.

Figure 1: ***(a)*** *Creativity happens at the frontier of understanding formed by existing concepts on the representation manifold learned from past experience.* ***(b)*** *The Creator-Appraiser framework, with a Creator C and an Appraiser A. A learns the concept c created by C in the inner loop, while C's goal in the outer loop is to make c* unfamiliar *to A at first but highly* learnable *or* memorable.

One path to building creative artifacts is through novel composition of existing parts. A mermaid is imagined as a human upper body fused with a fish tail; the verbal recipe makes the creation easily described and learned. Composition, however, is not the only path. Some concepts are perceived as monolithic: the color fields of Mark Rothko, the iconic blue of Yves Klein, or the splash patterns of Jackson Pollock are not naturally read as combinations of simpler parts. A theory of creativity must accommodate both, and the difference may lie in which layers of an information-processing system are engaged.

Finally, an account of conceptual creativity must address aesthetics, i.e. what makes a creation appealing to our senses. Beyond evolutionary preferences, a powerful computational handle is abstraction through compression. Many artistic forms feature complex symmetries: Islamic geometric patterns, Gothic architecture, and the repetition and structure of musical composition are all highly compressible, in the sense that a small fragment suffices to predict the whole. Compressibility, however, is not universal: it depends on what the observer already knows. Aesthetic judgment is therefore partly an "acquired taste"; Jazz can sound chaotic on first listen and rich once the listener has internalized its chord progressions and rhythmic motifs.

Together these examples sketch a picture of creativity at the *frontier of understanding*: existing concepts form a representation manifold built from past experience, and novel concepts lie just beyond it—past what an observer's current representations can immediately predict, but close enough that a small update can assimilate them (Fig. 1a). Schmidhuber (2010) formalized this as *compression progress*: a stimulus is interesting when it improves an observer's predictive model. What counts as creative, in this picture, is therefore observer-relative: a stimulus is creative *for* an observer whose existing representations cannot yet predict it but whose representations can quickly be updated to do so. We use adaptation, learning, and memorization interchangeably throughout, since our focus is the fast adaptation of an observer's memory to a single instance. As the title suggests, we hypothesize that creativity is the process of seeking concepts that are unfamiliar but memorable to the observer in question.

The rest of the paper develops this idea computationally by connecting it to meta-learning: an inner loop in which the observer (later formalized as an *Appraiser*) tries to assimilate the creative input, and an outer loop that refines the input itself so that it begins as unfamiliar and ends as memorable to that same observer after just a few internal updates.

2 Creator-Appraiser Meta-Learning

We define an *observer* as a *predictive system*: an encoder E that maps an input x to a representation $z = E(x)$, and a head H that emits a prediction whose discrepancy from x we write as $\mathcal{L}(H, E; x)$. The head can take many forms: a decoder that reconstructs x, a next-token predictor over a sequence, or a classifier that decides whether z belongs to x versus a set of reference stimuli. We say the observer *understands* x when the prediction error $\mathcal{L}(H, E; x)$ is small. This setup subsumes the predictive-modeling view of perception (Rumelhart et al., 1986; Hinton et al., 2006; Kingma and Welling, 2013; Vincent et al., 2008; Devlin et al., 2019; He et al., 2022; Bengio et al., 2003; Assran et al., 2023).

A creator-relevant observer does not stay frozen. When faced with a novel stimulus that its current representations fail to predict, an adaptive observer updates a subset of its parameters to accommodate it, and the creator gets useful feedback only because of this update. Concretely, novelty is the initial prediction error, and aesthetics is the rate at which a brief adaptation reduces that error. The value of a stimulus is then the size of this drop. We assign this adaptive observer to the role of an *Appraiser*: an observer that not only predicts but also reports its own learning progress back to the creator. We now formalize this in a bilevel optimization between a Creator that generates and an Appraiser that adapts.

The two-player formulation. We propose conceptual creation as a bilevel optimization between a Creator (C) that generates candidate concepts and an Appraiser (A) that adapts to them.

The Appraiser A is a predictive system parameterized by a pair of weights (θ, ϕ): θ are slow weights that hold A's prior over its inputs, fixed throughout creation; ϕ are fast weights that adapt within the inner loop. Given a candidate c from C, A runs an inner-loop optimization to fit c, minimizing its prediction error on c alone:

$$\min_{\phi} \mathcal{L}_A(\phi, \theta; c), \tag{1}$$

where $\mathcal{L}_A$ is the predictive loss of A, instantiated as reconstruction error, classification error, or any other discrepancy a predictive head produces on c. The improvement in this loss after T inner-loop steps, $\mathcal{L}_A(\phi_0, \theta; c) - \mathcal{L}_A(\phi_T, \theta; c)$, is the reward A reports back to C. C's objective is therefore to produce a c for which A's learning progress is large:

$$\min_{c} \mathcal{L}_C(c) = \min_{c} \mathcal{L}_A(\phi_T, \theta; c) - \mathcal{L}_A(\phi_0, \theta; c), \tag{2}$$

where ϕ_0 are the initial fast weights and ϕ_T are the weights after T inner-loop steps; C minimizes this composite objective with respect to its own parameters or sample space. Because high-reward c are not unique, we expect C to explore multiple trajectories rather than collapse to a single optimum. To ensure every candidate is evaluated under a fresh prior, the Appraiser's fast weights ϕ are reset to a fixed initialization ϕ_0 before each C iteration, so A encounters every c as a first impression. The full bilevel meta-gradient procedure is summarized in Algorithm 1 and illustrated in Figure 1b.

In our experiments, the outer iteration is run once per denoising step of a diffusion Creator, with $\tilde{c}$ being the predicted clean output $\hat{x}_0^{(t)}$ at step t and ξ being either the noise latent (MNIST) or the text embedding (natural images). The same template applies to any iterative or one-shot generator with a differentiable parameter ξ.

2.1 Appraiser instantiations

The framework is agnostic to the choice of Appraiser. We describe two concrete instantiations that we use in our experiments (Figure 2): an Autoencoder Appraiser (used for the MNIST experiments) and a CLIP Appraiser (used for the natural image experiments).

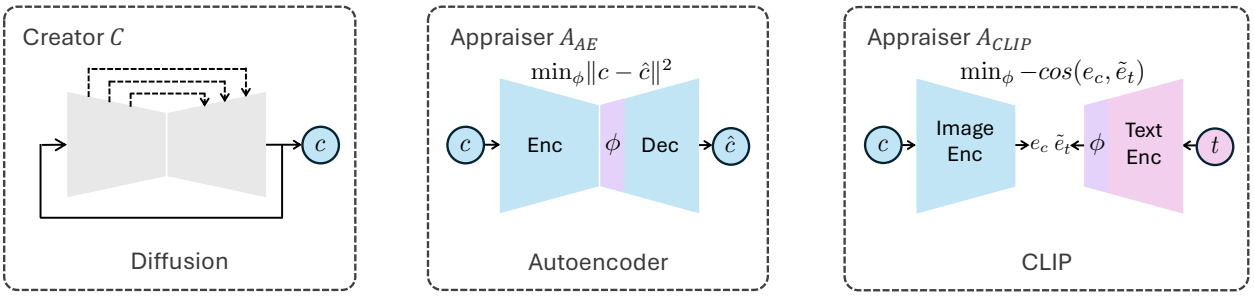


Figure 2: *Implementation of the Creator-Appraiser setup in the experiments. We illustrate both the autoencoder and CLIP appraisers. ϕ denotes the adaptation parameter in the Appraiser for concept learning.*

Algorithm 1 Creator-Appraiser Meta-Learning (one outer iteration)

Require: Creator C with learnable variable ξ; Appraiser A with slow weights θ and fast-weight initialization ϕ_0; loss $\mathcal{L}_A$; inner steps T, inner LR α; outer LR η.

$\tilde{c} \leftarrow C(\xi)$ ▷ generate (intermediate or final) candidate
$\phi \leftarrow \phi_0$ ▷ reset Appraiser fast weights for first impression
$\ell_0 \leftarrow \mathcal{L}_A(\phi, \theta; \tilde{c})$
for $k = 1, \dots, T$ **do** ▷ inner-loop adaptation, differentiable in ξ
    $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}_A(\phi, \theta; \tilde{c})$
end for
$\ell_T \leftarrow \mathcal{L}_A(\phi, \theta; \tilde{c})$
$R \leftarrow \ell_0 - \ell_T$ ▷ learning-progress reward
$\xi \leftarrow \xi + \eta \nabla_\xi R$ ▷ meta-gradient update on Creator

2.1.1 Autoencoder Appraiser

The encoder E is frozen and pretrained (e.g. classification, self-supervised learning, etc.); the decoder H is a lightweight network pretrained to reconstruct training images from E's features. The inner loop fine-tunes the decoder on the single candidate image c with reconstruction loss (MSE), and the reward is the loss improvement after K steps. The Appraiser thus rewards inputs the decoder can quickly assimilate from its existing representations, and gives near-zero reward on noise-like inputs that the decoder cannot fit at all.

Reward shaping. We refer to $\mathcal{L}_{A,0}$ as the *unfamiliarity* of c (how surprising it is to a naive observer) and $\mathcal{L}_{A,T}$ as the *residual unfamiliarity* (how much remains unpredictable after adaptation). For the autoencoder Appraiser, the raw improvement $\mathcal{L}_{A,0} - \mathcal{L}_{A,T}$ over-rewards high-noise inputs (large initial reconstruction loss to recover from) even when no meaningful structure has emerged. We therefore multiply the improvement by two Gaussian-shaped gates that constrain $\mathcal{L}_{A,0}$ and $\mathcal{L}_{A,T}$ to a "learnable yet unfamiliar" region:

$$R = (\mathcal{L}_{A,0} - \mathcal{L}_{A,T}) \cdot g_0(\mathcal{L}_{A,0}) \cdot g_T(\mathcal{L}_{A,T}), \tag{3}$$

$$g_0(\mathcal{L}_{A,0}) = \exp\left(-\frac{(\mathcal{L}_{A,0} - \mu)^2}{2\sigma^2}\right), \qquad g_T(\mathcal{L}_{A,T}) = \exp\left(-\frac{\mathcal{L}_{A,T}}{\tau}\right). \tag{4}$$

The *complexity gate* g_0 keeps the initial prediction error in a target range around μ (avoiding both familiar low-$\mathcal{L}_{A,0}$ inputs and noise-like high-$\mathcal{L}_{A,0}$ inputs). The *success gate* g_T rewards low final error, ensuring the decoder actually fits the input. We show the effect of applying the gates in Figure 4.

2.1.2 CLIP Appraiser

The Autoencoder Appraiser ties learnability to a domain-specific encoder–decoder pair pretrained on the target distribution. This is a clean fit for purely visual settings like MNIST, where the candidate is a small image and a single-modality reconstruction loss reads as concept fidelity. Natural image generation is different in two ways. First, many pretrained creative models are *text-conditioned*: the prompt names the concept and the Appraiser must judge whether the candidate fits that concept, so the learnability signal needs to live in a joint vision–language space. Second, even setting that aside, pretraining a high-fidelity natural-image decoder is expensive, and the reconstruction loss would be dominated by pixel-level texture rather than concept-level fit. We therefore instantiate the Appraiser on top of a frozen vision-language encoder pair, where the inner loop adapts a small text-side projection to match the generated image, providing a concept-level learnability signal that is naturally text-conditioned.

We use a contrastive image-text model CLIP (Radford et al., 2021) with both image and text encoders frozen, projecting into a shared embedding space where matching image-text pairs have high cosine similarity. The inner loop fine-tunes a low-rank adapter (Hu et al., 2022) of rank r on the text projection layer, $W'_{\text{text}} = W_{\text{text}} + AB$, with A initialized to zero. The K-step inner loop updates A and B to maximize $\cos(e_{\text{img}}, \tilde{e}_{\text{text}})$, where $\tilde{e}_{\text{text}}$ is the LoRA-adapted text feature; the reward is the similarity improvement $R = \text{sim}_T - \text{sim}_0$. This is the natural CLIP analog of the Autoencoder Appraiser: both fine-tune a small predictor against a target representation, asking how well a small adaptation can match the candidate. Rank r restricts adaptation capacity: at small r, the inner loop cannot trivially memorize any image's direction in K steps. With the LoRA adapter, the similarity-improvement reward R alone is sufficient without the need for additional reward shaping.

3 Experiments

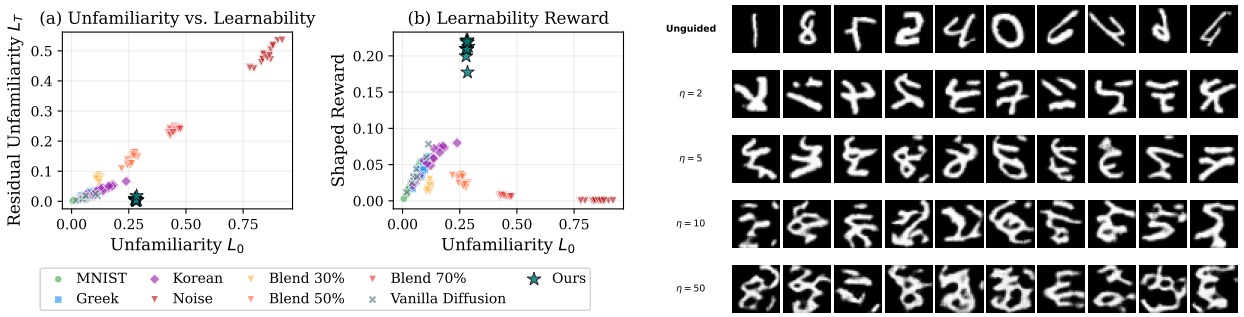
We evaluate our framework in two settings: (1) an MNIST setting that validates the core learnability signal and the unfamiliarity–learnability frontier with the Autoencoder Appraiser, and (2) a natural image setting using Stable Diffusion with the CLIP Appraiser.

3.1 MNIST

Setup. The Appraiser uses a LeNet-style encoder (Conv→ReLU→Pool, two layers, 512-dimensional feature space) and a transposed-convolutional decoder that reconstructs 32×32 grayscale images from the feature bottleneck. Both are pretrained on MNIST: the encoder with classification loss, the decoder with pixel-level MSE reconstruction loss. The Creator is a pretrained DDPM diffusion model (Ho et al., 2020) that generates MNIST-like images. For brevity in figures and text we write $L_0 = \mathcal{L}_{A,0}$ and $L_T = \mathcal{L}_{A,T}$.

Learnability signal validation. Figure 3a probes whether the shaped reward correctly orders stimuli that span the unfamiliarity–learnability space: MNIST test digits (familiar, structured), Greek and Korean characters from Omniglot (unfamiliar, structured), Gaussian noise (unfamiliar, unstructured), and noise-blended MNIST digits at 30%, 50%, and 70% noise ratios. The blended images have similar L_0 to structured Omniglot scripts but lack coherent structure. Panel (a) shows the (L_0, L_T) plane: noise in the top-right corner (high L_0, high L_T—unfamiliar but unlearnable), familiar digits in the bottom-left, and our guided generations in the bottom-middle (moderately unfamiliar yet quickly learnable). Panel (b) ranks categories by shaped reward, separating Korean and Greek from same-L_0 noise blends and pushing noise to near-zero.

Why the improvement term matters. A natural objection is that the shaped reward reduces to gated unfamiliarity—if the complexity gate already filters by L_0, why also include the improvement term? Fig-



(a) Unfamiliarity–learnability space.

(b) Guided samples at increasing η.

Figure 3: ***(a)*** *Our guided images occupy the bottom-middle of the (L_0, L_T) plane—moderately unfamiliar yet highly learnable; noise lies in the top-right corner, and our shaped reward significantly exceeds the best reference category (Korean).* ***(b)*** *As guidance scale η increases, generated outputs become progressively more novel and complex while retaining stroke-like structure.*

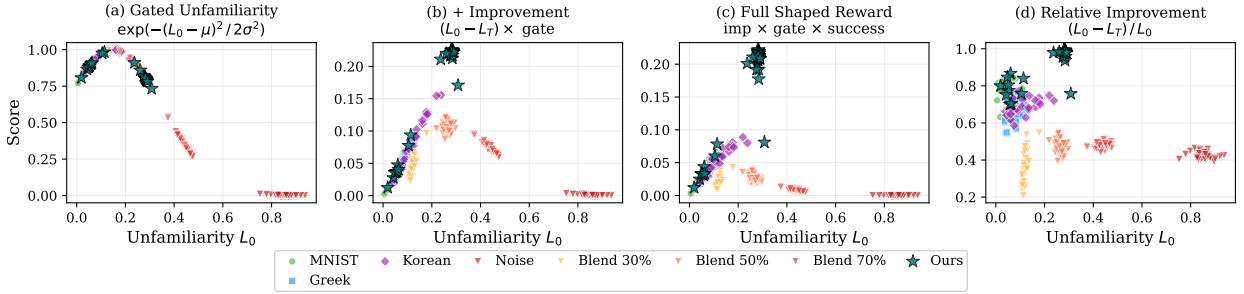


Figure 4: *Decomposition of the shaped reward.* ***(a)*** *The complexity gate alone cannot distinguish structured from degraded stimuli at similar L_0.* ***(b)*** *Adding the improvement term separates structured scripts from noise blends.* ***(c)*** *The full shaped reward.* ***(d)*** *Relative improvement inverts the desired ranking, rewarding familiarity over structured novelty.*

ure 4 ablates each reward component. Panel (a), complexity gate alone, cannot distinguish structured Korean characters from noise-blended digits at the same L_0: both pass the gate. Panel (b), complexity gate + improvement term, separates them, since Korean characters admit a much larger reconstruction-loss drop than noise blends with comparable initial error. Panel (c) shows the full shaped reward, which combines (b) with the success gate g_T to additionally suppress inputs the decoder cannot fit at all. Panel (d) is a control: replacing the absolute improvement with the *relative* improvement $(L_0 - L_T)/L_0$ inverts the ranking, rewarding familiar inputs (small L_0, small relative error) over structured-novel inputs.

Layer-specific learnability. We further ablate which decoder layers are fine-tuned during the inner loop (Appendix Figure 13). Fine-tuning only the *first* (high-level, semantic) decoder layer yields improvement comparable to fine-tuning the full decoder, while fine-tuning only the *last* (low-level, textural) layer yields near-zero improvement across all categories. This confirms that learnability is driven by adaptation at the semantic feature level, not by low-level pattern fitting.

Guidance scale and ideal initial loss. The two main hyperparameters of the guided sampler, the guidance scale η and the complexity-gate target μ, control complementary axes of the output (Appendix Figures 9 and 11). η controls how strongly the appraiser's gradient pushes the latent updates: as η grows,

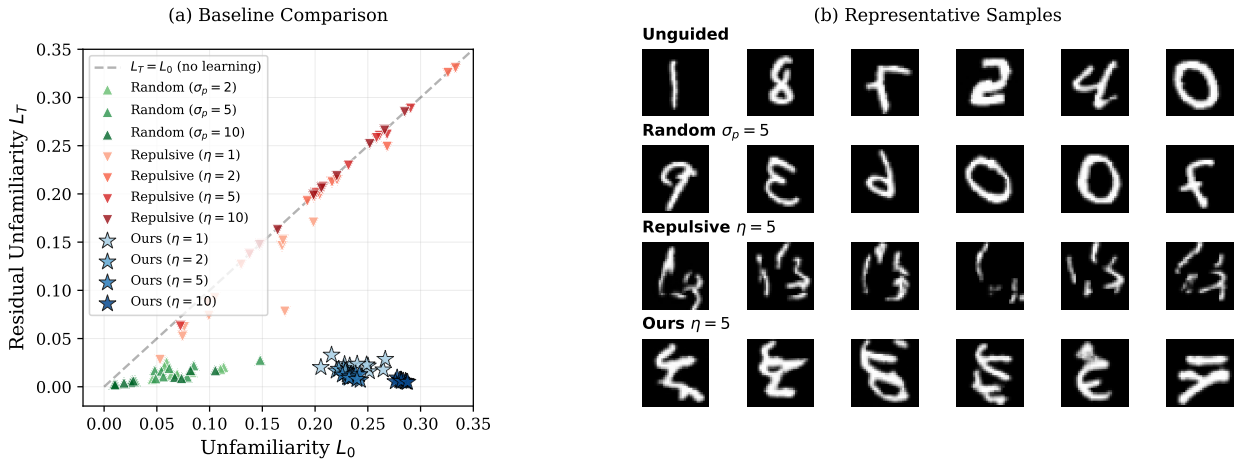


Figure 5: *Baseline comparison.* ***(a)*** *Each point is one generated image in the* (L_0, L_T) *plane; the dashed diagonal marks* $L_T = L_0$ *(no learning). Color gradients show increasing* η *(or* σ_p*) within each method. Random perturbation stays familiar (near origin). Repulsive guidance increases unfamiliarity but rides the diagonal where no learning occurs. Only our method achieves the bottom-right: unfamiliar* and *learnable.* ***(b)*** *Representative samples at* $\eta = 5$ *(*$\sigma_p = 5$ *for random).*

samples become progressively more learnable but also visually more complex, with the higher-η outputs reading as busier and less aesthetically restrained than the lower-η ones. μ instead shifts the unfamiliarity target: lower μ pulls outputs closer to the training distribution, higher μ pushes them further from it. The two together act as independent dials on *how unfamiliar* and *how learnable* the generated images are.

Comparison with baselines. We compare against three baselines that each isolate a component: **Vanilla DDPM**; **noise perturbation** of scale $\sigma_p \in \{2, 5, 10, 20\}$ added to the latent at each step (tests whether unstructured perturbation alone is sufficient); and **repulsive guidance** (Lu et al., 2024) with scale $\eta \in \{1, 2, 5, 10, 20, 50, 100\}$ pushing the predicted latent away from class prototypes (deviation without learnability). Figure 5(a) plots all methods in the (L_0, L_T) plane: noise stays near the origin (the DDPM absorbs the perturbation, so neither L_0 nor L_T moves much), repulsive guidance rides the diagonal $L_T \approx L_0$ (deviation without structure, the appraiser cannot fit these outputs any better than it could a noise input of the same L_0), and only our method reaches the bottom-right, below the no-learning diagonal at high unfamiliarity. The visual difference matches the quantitative ranking (Fig. 5b): noise-perturbed samples are visually indistinguishable from unguided digits; repulsive-guided samples are clearly disrupted but unstructured; ours produce coherent stroke-like glyphs with consistent line weight. Additional samples from each method are in Appendix Figure 8.

3.2 Natural Images

Setup. We instantiate the framework on natural image generation using the CLIP Appraiser (Section 2.1.2) paired with a Stable Diffusion v1.5 Creator (Rombach et al., 2022). Unlike the MNIST setting, where the Creator's learnable variable is the latent noise z and the meta-gradient flows back to z at each denoising step, here the learnable variable is the *text embedding*: the prompt is encoded once, the encoded representation becomes a learnable parameter, and the meta-gradient updates it at each denoising step. This design choice puts the search in semantic space rather than pixel/latent space, which we found necessary for the appraiser's reward to track concept-level fit rather than pixel texture. We evaluate two types of

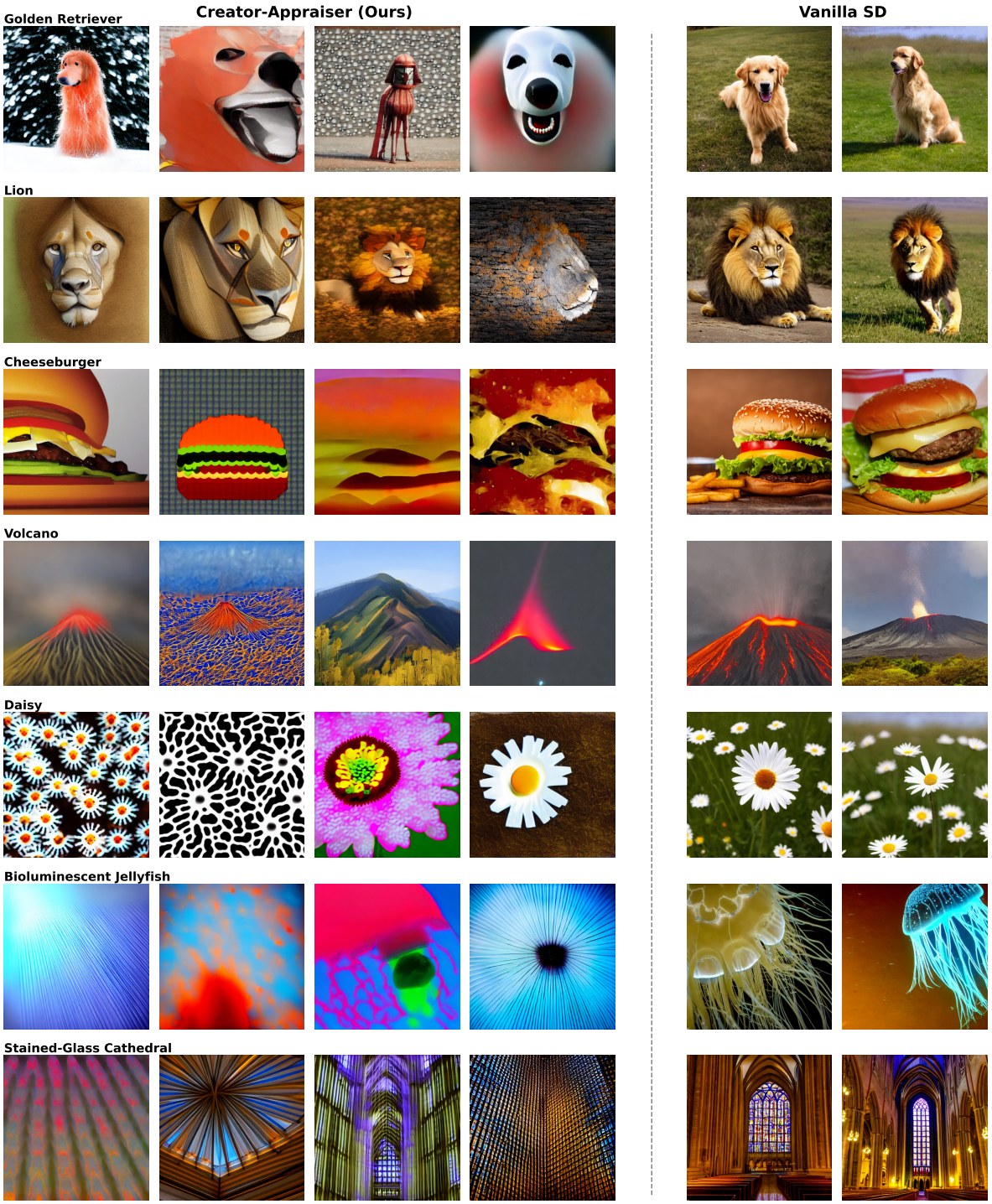


Figure 6: *Curated outputs from the CLIP Appraiser at LoRA rank $r \in \{8, 32\}$, $\eta = 10$. Each row shows four samples from our method (left) and two vanilla SD samples (right) for the same prompt; rows 1–5 are ImageNet class-conditioned, rows 6–7 are open-ended text prompts. The meta-gradient produces visually distinct stylistic interpretations within each prompt.*

prompts: **(i) class-conditioned prompts** of the form "a photo of a {class}" for ImageNet classes, and **(ii) open-ended prompts** drawn from a curated set of stylistic descriptions (e.g., "stained-glass cathedral interior at night"). For each prompt we sweep LoRA rank $r \in \{4, 8, 16, 32\}$ at outer-loop guidance $\eta = 10$.

Measures. We report three quantitative measures. (i) $\mathbf{sim}_0$: the cosine similarity between the generated image's CLIP embedding and the prompt's text embedding under the *frozen* CLIP model; this is also the inner-loop initial similarity. (ii) $\Delta\mathbf{sim} = \mathbf{sim}_T - \mathbf{sim}_0$: the similarity improvement after the K-step LoRA inner loop; this is the inner-loop reward, and what the meta-gradient optimizes. Because sim_T uses the adapted CLIP and sim_0 uses the frozen one, the two are independent. (iii) **VLM relevance score** using a Qwen2.5-VL-7B (Bai et al., 2025) on a 0–10 scale, judging whether the prompt's concept is present in the image; the VLM pass rate (fraction with score ≥ 5) is a stricter, CLIP-independent on-prompt check, with details in Appendix B.

Results. Figure 6 shows representative picks; the vanilla SD baselines on the right are uniformly photorealistic and on-prompt. The picks veer into concepts that the text prompt never asked for: a lion in autumn-leaf form and a metamorphosis-style lion side face; multiple golden retrievers in novel coloration, body shape, and proportion; a cheeseburger rendered as a sunset landscape; a daisy reduced to a black-and-white interior–exterior abstraction where the petals dissolve into ink strokes; another daisy with an egg-yolk center ringed by tissue-like rectangles; a pastel oil-painted volcano; magenta-colored bioluminescent jellyfish. None of these are in the prompt—the input string was simply "a photo of a daisy" or "a photo of a lion"—and the diffusion weights are not fine-tuned. The entire creative deviation is discovered inside the latent prompt embedding by the meta-gradient acting on the inner-loop learning-progress signal.

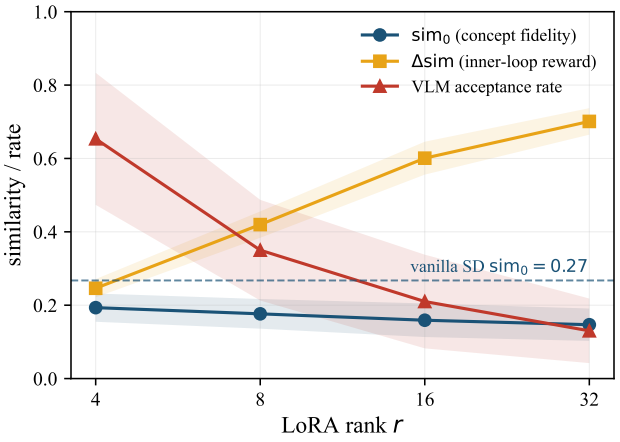


Figure 7: *Effect of LoRA rank on the natural-image setup. As r grows, the inner-loop reward Δsim rises sharply while frozen-CLIP sim_0 drifts down from the vanilla baseline; the Qwen2.5-VL relevance acceptance rate (fraction of samples judged on-prompt) drops from 65% at r=4 to 12% at r=32. The two latter curves trace the capacity-vs-fidelity trade-off: more inner-loop capacity buys a stronger reward signal at the cost of off-class drift. Per-prompt breakdown in Table 2.*

Capacity-vs-fidelity trade-off. Figure 7 traces this trade-off across the four LoRA ranks we sweep: vanilla SD has $\text{sim}_0 \approx 0.27$ with no adapter to apply; at r=4 our method already drops sim_0 to 0.19 with $\Delta\text{sim} \approx 0.25$, and as rank grows Δsim rises monotonically to 0.70 at r=32 while sim_0 continues to drift down to 0.15. The cost shows up in the VLM-relevance acceptance rate, which falls from ≈ 65% at r=4 to ≈ 12% at r=32: lower rank acts as the capacity gate, the inner LoRA cannot memorize off-manifold images and the meta-gradient is implicitly constrained to stay near the prompt's class; higher rank lifts that constraint and occasionally pushes embeddings into stylistically dramatic but off-class regions.

Reachability beyond vanilla. Are our creative outputs simply re-discoveries of samples that vanilla Stable Diffusion would also produce, given enough draws? To answer this, for each of the 20 prompts we test we draw $N = 1000$ vanilla samples to form a sample set V, embed both vanilla and Creator-Appraiser samples with a frozen DINOv2 ViT-L/14 (Oquab et al., 2024), and compute nearest-neighbor cosine distance.

To exclude off-class cases, we restrict the CA side to samples that pass a Qwen2.5-VL relevance check at ≥ 5 on a 0–10 scale. On this comparison, our r=8 outputs achieve a median 5× (range 2–20×) farther from any vanilla sample than vanilla samples land from each other; the CA sample distribution is also $\sim$ 3.7× wider internally than vanilla, ruling out a single shifted mode. The gap grows monotonically with LoRA rank: median $\bar{d}_V^C$ rises from 0.42 at r=4 to 0.70 at r=32 while the intra-vanilla baseline holds at $\sim$ 0.11. This result shows that the meta-gradient discovers regions of the diffusion model's output space that vanilla sampling does not reach at $N = 1000$. Per-prompt numbers and the nearest neighbor sample figure are in Appendix C.

Visualization on an automatic top-K selection with VLM filtering (no manual curation) is in Appendix B, and more detailed visual analysis is in Appendix C.

4 Related Work

Computational creativity. A persistent question is whether machines can create at all; Ada Lovelace famously argued that Babbage's Analytical Engine "has no pretensions whatever to *originate* anything" (Lovelace, 1843). We take the position that creativity is best defined operationally, by its *effect on an observer* rather than by the creator's experiential phenomenology (Nanay, 2014); on that definition, machines can create, and the question becomes how to build them. The broader field of Computational Creativity has modeled creative behavior in machines for decades (see survey (Franceschelli and Musolesi, 2024)). Boden (2004)'s distinction between *exploratory* (within an existing space) and *transformational* (reshaping the space itself) creativity is the standard reference, though in our framework the two interact: each new creation both probes the observer's current representations and updates them.

Schmidhuber (2010) proposed a formal theory of creativity based on compression progress: a stimulus is "interesting" to an observer when it improves the observer's predictive model. The key difference in our work is that the Creator is *aware* of the Appraiser's inner-loop learning progress: it is updated by meta-gradient through the Appraiser's adaptation, not just by a scalar reward. Realizing this in a modern vision generative model requires differentiating through diffusion sampling and through an inner loop on a CLIP-LoRA Appraiser. Follow-up work (e.g. POWERPLAY (Srivastava et al., 2013)) aimed at RL task completion rather than visual creation. Our reward shaping also resonates with Berlyne (1960)'s arousal theory in psychology, where aesthetic pleasure peaks at intermediate novelty and complexity, neither too simple nor too overwhelming.

Information theory and bounded learnability. Classical information measures abstract away the bounded observer: Shannon entropy treats pure noise as maximally informative, while Kolmogorov complexity (Kolmogorov, 1965) assumes an unbounded observer for whom any compressible structure is free. Neither captures structure a *bounded* learner can extract with a small adaptation. The Minimum Description Length principle (Rissanen, 1978) formalizes learning as two-part coding, and Bennett (1988)'s *logical depth* ties meaning to the computational effort required to produce a structure. Most closely related, Finzi et al. (2026) recently introduced *epiplexity*: the length of the shortest program achieving the optimal two-part code at a given compute budget. Our Creator-Appraiser reward $\mathcal{L}_{A,0} - \mathcal{L}_{A,T}$ measures observer-relative bounded learnability via a *specific* Appraiser rather than a search over learners. Crucially, epiplexity is a measurement on existing data; our framework adds a generative side, with the Creator producing new artifacts by optimizing the Appraiser's reward via meta-learning gradients.

Deep generative models and creativity. Modern deep generative models—Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), particularly StyleGAN (Karras et al., 2019), and more recently

diffusion models (Ho et al., 2020)—have become the dominant paradigm for high-fidelity image generation by modeling the underlying data manifold. Creative Adversarial Networks (Elgammal et al., 2017) extended GANs by training the generator to deviate from learned artistic styles while remaining aesthetically plausible. Earlier work (Gatys et al., 2016) hinted at a layered separation of content and style in deep features. However, the apparent creativity of these models is, by default, largely a consequence of recombining concepts and styles already present in the training distribution (Richardson et al., 2024; Feng et al., 2025), and obtaining outputs further from typical sampling generally requires careful prompt engineering.

Several recent works have proposed explicit mechanisms for creative generation in diffusion models. Lu et al. (2024) guide diffusion away from existing training examples using a propulsive energy term that repels generated samples from specific reference images. Han et al. (2025) enhance creativity by amplifying low-frequency components in early diffusion layers. Golan et al. (2025) use vision-language models to adaptively construct negative prompts that steer generation toward unconventional outputs. While these methods differ in mechanism, they share a common strategy: encouraging the generator to deviate from its typical output distribution. Our approach introduces an orthogonal objective. Rather than defining creativity through deviation alone, we optimize for *learnability*—the generated output must be not only unfamiliar to the observer but also structured enough to be rapidly learned. Deviation-based methods can produce novel but incoherent outputs; learnability provides a principled constraint that selects for structured novelty.

Adversarial, actor-critic, and meta-learning frameworks. The Creator-Appraiser structure resembles the generator-discriminator dynamic in GANs (Goodfellow et al., 2014) and the actor-critic framework in RL (Sutton and Barto, 2018), with two key differences. Unlike a GAN discriminator, which trains toward a fixed decision boundary at equilibrium, the Appraiser is a *dynamic learner* re-initialized for each creation, and the Creator is rewarded not for fooling it but for the inner-loop *learning progress* on that artifact. The actor-critic analogy is closer: like a critic the Appraiser scores artifacts, but its "value" is the *learning progress* the Appraiser itself makes when adapting to the just-produced artifact, not a long-horizon return, in the spirit of intrinsic-motivation rewards based on prediction improvement (Schmidhuber, 2010; Pathak et al., 2017). There is no environment, no temporal credit assignment, and no shared parameters; the Appraiser is reset between creations rather than slowly tracking a moving target. The inner-loop adaptation draws on meta-learning (Finn et al., 2017), and the text-embedding optimization relates to Textual Inversion (Gal et al., 2023), though with a learnability rather than reconstruction objective.

5 Conclusion

We have proposed a Creator-Appraiser framework that casts creativity as the production of artifacts initially unfamiliar yet rapidly learnable to an adaptive observer, and showed that a meta-gradient through a brief inner-loop adaptation reliably produces diffusion outputs substantially different from repeated sampling at a fixed prompt. Our framing is currently limited to the *perceptual* value but not functional value, and the criterion is plausibly necessary but not sufficient for what humans call creativity. The Appraiser is also memoryless across creations; an observer with longer-range memory could in principle mark a stimulus as "learnable in the context of having recently seen related stimuli," opening room for context-dependent and curriculum-driven creativity. Finally, while we instantiate the framework on image diffusion, the Creator-Appraiser structure transfers to any generator paired with an adaptive predictor; extending it to language, music, or interactive design can be natural future directions.

Acknowledgment

This work was supported in part by Visko AI, a Google TPU Award, and the Institute of Information & Communications Technology Planning Evaluation (IITP) under grant RS-2024-00469482, funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research. The compute is supported by the NYU High Performance Computing resources, services, and staff expertise.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., et al. (2025). Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3(Feb):1137–1155.

Bennett, C. H. (1988). Logical depth and physical complexity. In Herken, R., editor, *The Universal Turing Machine: A Half-Century Survey*, pages 227–257. Oxford University Press.

Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity*. McGraw-Hill.

Boden, M. A. (2004). *The Creative Mind: Myths and Mechanisms*. Routledge.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. (2017). Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611.

Feng, F., Xie, Y., Yang, X., Wang, J., and Geng, X. (2025). Distribution-conditional generation: From class distribution to creative generation. *arXiv preprint arXiv:2505.03667*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*. PMLR.

Finzi, M., Qiu, S., Jiang, Y., Izmailov, P., Kolter, J. Z., and Wilson, A. G. (2026). From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv preprint arXiv:2601.03220*.

Franceschelli, G. and Musolesi, M. (2024). Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Golan, S., Nitzan, Y., Wu, Z., and Patashnik, O. (2025). Vlm-guided adaptive negative prompting for creative generation. *arXiv preprint arXiv:2510.10715*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*.

Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.-W., and Choi, J. (2022). Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*.

Han, J., Kwon, D., Lee, G., Kim, J., and Choi, J. (2025). Enhancing creative generation on stable diffusion-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):3–11.

Lake, B., Salakhutdinov, R., Gross, J., and Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*.

Lovelace, A. (1843). Notes upon L.F. Menabrea's "sketch of the analytical engine invented by Charles Babbage". *Taylor's Scientific Memoirs*, 3:666–731.

Lu, J., Teehan, R., and Ren, M. (2024). Procreate, don't reproduce! propulsive energy diffusion for creative generation. In *Computer Vision - ECCV 2024*. Springer.

Nanay, B. (2014). An experiential account of creativity. In Paul, E. S. and Kaufman, S. B., editors, *The Philosophy of Creativity: New Essays*. Oxford University Press.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Richardson, E., Goldberg, K., Alaluf, Y., and Cohen-Or, D. (2024). Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics*, 43(3):1–14.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*.

Srivastava, R. K., Steunebrink, B. R., and Schmidhuber, J. (2013). First experiments with PowerPlay. *Neural Networks*, 41:205–212.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1096–1103.

Appendix

A Additional Experiments

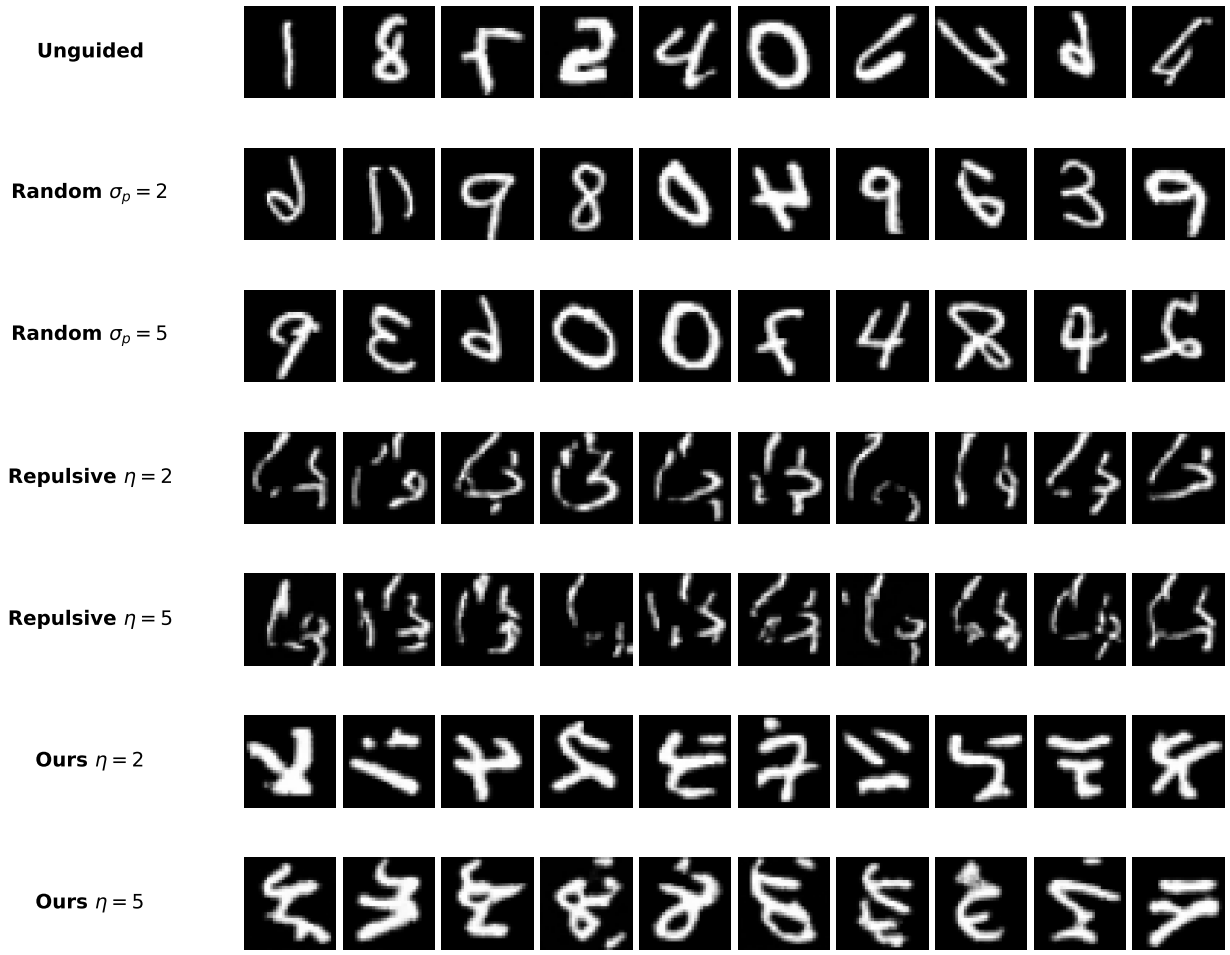


Figure 8: *Representative images from each method.* ***Unguided****: standard DDPM digits.* ***Random perturbation*** *($\sigma_p = 2, 5$): outputs are nearly identical to unguided—the diffusion process absorbs the noise.* ***Repulsive guidance*** *($\eta = 2, 5, 10$): images become progressively corrupted; higher repulsion produces more unfamiliar but incoherent outputs.* ***Ours*** *($\eta = 2, 5$; $\mu = 0.08$): novel, glyph-like patterns with coherent stroke structure.*

B Automatic Top-K Pipeline and Outputs Across More Prompts

The samples in the main-text Figure 6 were manually curated. To complement them with an evidence of representativeness, we describe a fully automatic curation pipeline below and apply it to a new set of prompts not shown in the main text (Figure 14).

Generation. For each new prompt, we run the same Creator-Appraiser pipeline as in Section 3.2: a Stable Diffusion v1.5 Creator with the prompt embedding as the learnable variable, a frozen CLIP ViT-L/14 image encoder, and a LoRA inner loop on CLIP's text projection. We sweep LoRA rank $r \in \{4, 8\}$ and produce

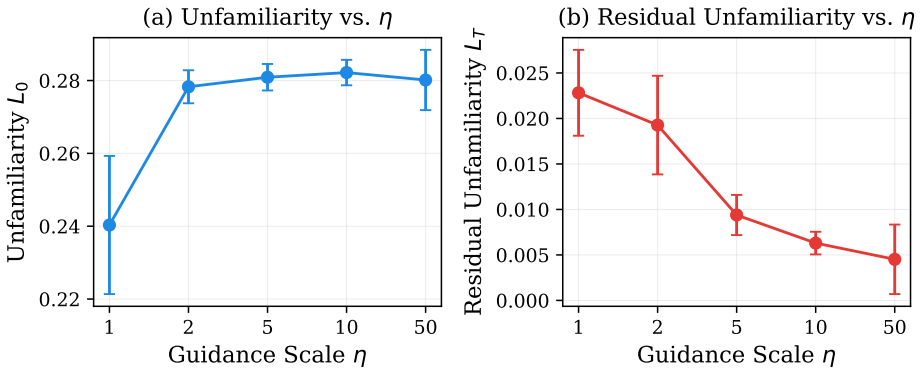


Figure 9: *Effect of guidance scale η on generated image statistics ($\mu = 0.15$).* ***(a)*** *Unfamiliarity L_0 saturates quickly beyond $\eta = 2$.* ***(b)*** *Residual unfamiliarity L_T drops monotonically—higher η produces more learnable outputs.*

100 samples per (prompt, rank) cell at outer-loop guidance $\eta = 10$, $K = 10$ inner-loop steps with learning rate 10^{-3}, and 50 denoising steps.

Per-image scoring. Every generated image is scored along three axes:

- sim_0, the frozen-CLIP image-text cosine similarity with the prompt (no LoRA). This is the inner-loop initial similarity.
- $\Delta\text{sim} = \text{sim}_T - \text{sim}_0$, the similarity improvement after a fresh K-step LoRA inner loop on the saved image. This re-scores the image against the same predictor used during generation.
- A Qwen2.5-VL-7B-Instruct relevance score (Bai et al., 2025) on a 0–10 scale, prompted to judge *whether the prompt's concept is recognizable in the image*, explicitly instructed not to lower the score for stylistic departure. This signal is independent of CLIP; the two encoders share no parameters.

Gating and ranking. We pass an image through the gate if $\text{sim}_0 \geq 0.12$ AND the VLM relevance is ≥ 5. The sim_0 gate only rejects clearly-broken outputs, and the VLM gate is a more meaningful relevance check. Survivors are ranked by Δsim in descending order; the top-8 are kept as candidate picks.

Sampling for the figure. For each prompt we randomly sample 6 of the top-8 with a fixed seed. The sampling step gives some within-prompt diversity (avoiding the degenerate case where all top-K are subtle variations of the single highest-Δsim image) without re-introducing manual curation.

C Reachability: DINOv2 Distance from the Vanilla Samples

We measure this with $N = 1000$ vanilla samples per prompt; embed both vanilla and CA outputs with a frozen DINOv2 ViT-L/14 (Oquab et al., 2024), L_2-normalize, and report nearest-neighbor cosine distances, mirroring the per-sample rarity construction of Han et al. (2022). For a CA sample x, $d_V(x) = \min_{x' \in V} \|f(x) - f(x')\|$ measures how far x is from the vanilla set V; for a vanilla sample x', $d_V(x') = \min_{x'' \in V \setminus \{x'\}} \|f(x') - f(x'')\|$ is the typical distance within V.

Shown in Table 1, our r=8 outputs are a median 5× farther from any vanilla sample than vanilla samples vary from each other. Across all 20 prompts in our figures, this ratio ranges from 2× (neon samurai in rain) up to 20× (peacock). The reachability gap grows monotonically with LoRA rank: median $\bar{d}_V^C$ rises

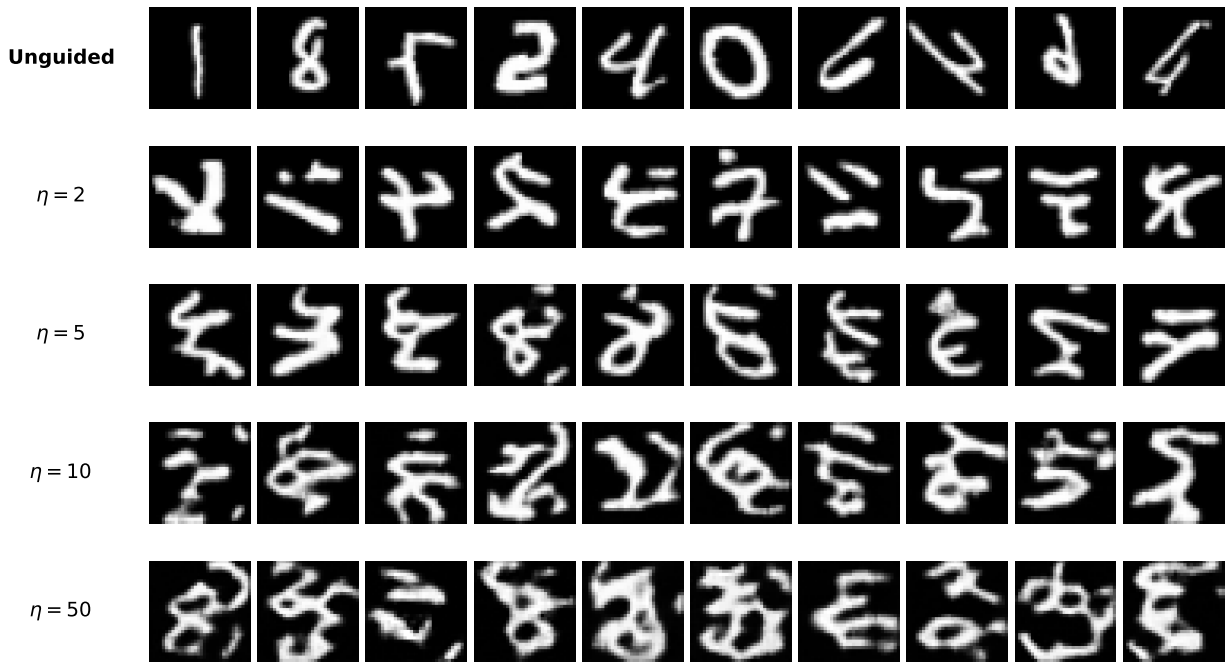


Figure 10: *Generated images at increasing guidance scales η, corresponding to the analysis in Figure 9. Unguided DDPM samples resemble MNIST digits; our guided outputs become progressively more novel while retaining stroke-like structure.*

from 0.42 at r=4 to 0.70 at r=32 while the intra-vanilla baseline holds at $\sim$ 0.11. The CA distribution is also $\sim$ 3.7× wider internally than the vanilla is from itself, ruling out the "shifted single mode" interpretation. Simply sampling more from a vanilla diffusion model does not reach the same outputs at $N = 1000$.

Figure 15 visualizes this directly: each picked CA sample is paired with its DINOv2-nearest vanilla counterpart from the 1000-sample set. In the rightmost pair of the Daisy row of Figure 15, our sample matches the vanilla counterpart's lighting, background tone, and even rough framing, but the flower head itself adopts a different geometry, with petals that read as ink-stroke abstractions rather than the photographic daisy that vanilla converges to.

Effect of LoRA rank. Figure 16 shows the same prompt rendered at $r \in \{4, 8, 16, 32\}$, with the top Best-of-N pick (scored by Qwen2.5-VL relevance $\geq$ 5, ranked by Δsim) shown per cell. The progression is monotone in two senses: visually, low-rank cells stay close to the photograph and high-rank cells push toward stylization, abstraction, or texture; quantitatively, $\bar{d}_V^C$ and the inner-loop reward Δsim both grow with rank (Table 1).

D Per-Prompt Natural Image Results

Table 2 reports the per-prompt frozen-CLIP similarity sim_0 and inner-loop reward Δsim for vanilla Stable Diffusion and our Creator-Appraiser at LoRA ranks $r \in \{4, 8, 16, 32\}$, with CA samples filtered by the Qwen2.5-VL relevance gate ($\geq$ 5). Vanilla sim_0 ranges from 0.23 to 0.34 across prompts (mean 0.27); on the same prompts our method drops sim_0 to 0.19 at r=4 and 0.15 at r=32 while the inner-loop reward Δsim rises from 0.25 to 0.70. The reachability metric for the same prompts and ranks is in Table 1.

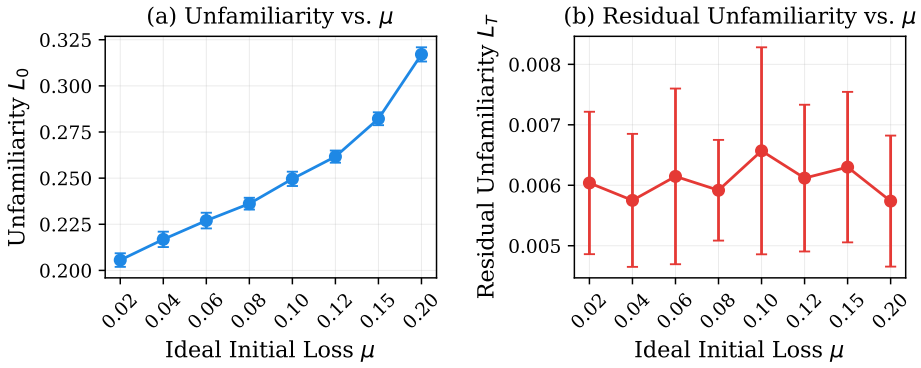


Figure 11: *Effect of ideal initial loss μ on generated image statistics ($\eta = 10$).* ***(a)*** *Unfamiliarity L_0 scales linearly with μ, confirming the complexity gate controls the unfamiliarity target as designed.* ***(b)*** *Residual unfamiliarity L_T remains flat (≈ 0.006) regardless of μ—learnability is consistently high.*

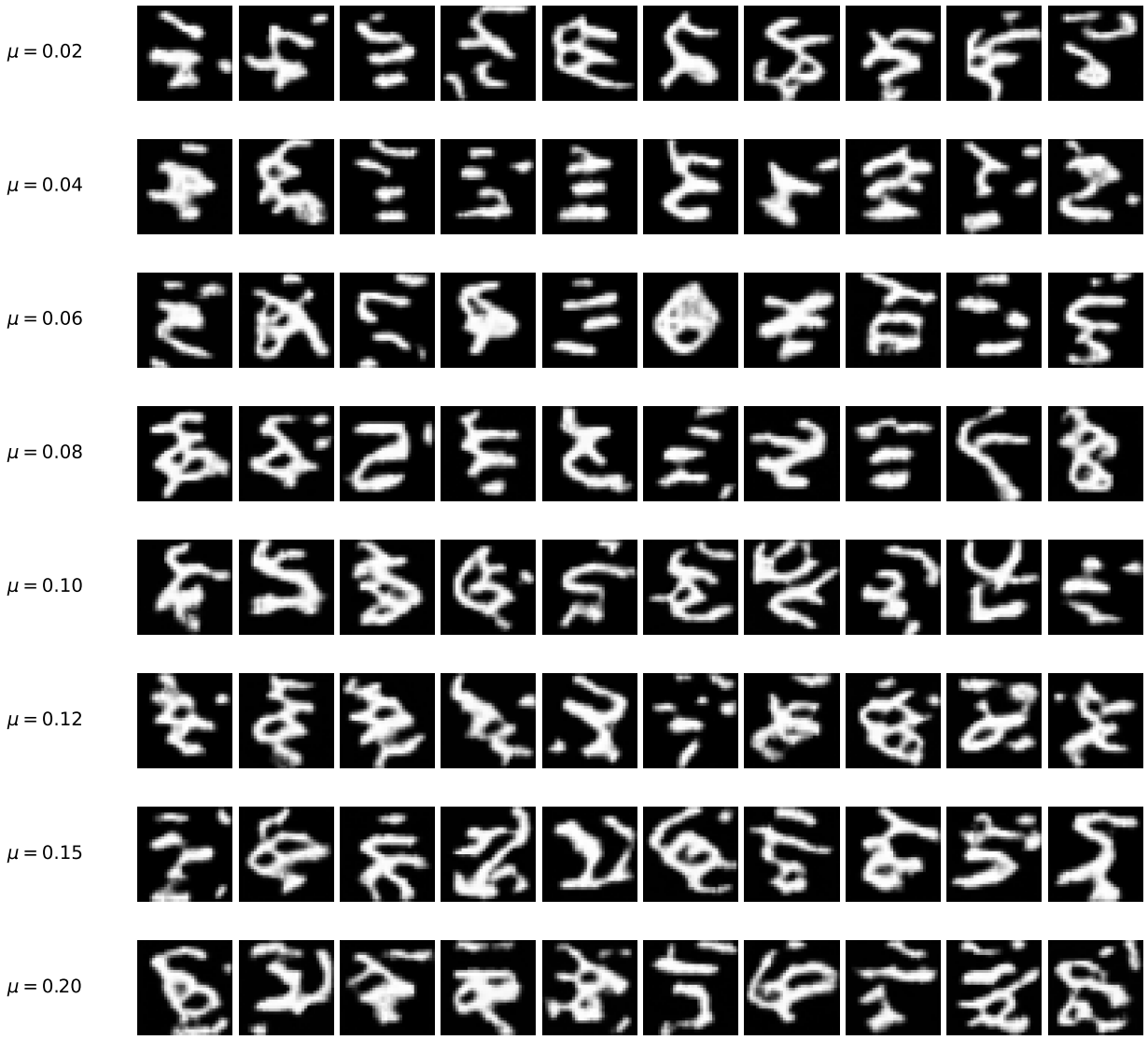


Figure 12: *Generated images across ideal initial loss μ values ($\eta = 10$), corresponding to the analysis in Figure 11. Lower μ produces thinner, more glyph-like forms; higher μ encourages thicker strokes and greater departure from the training distribution.*

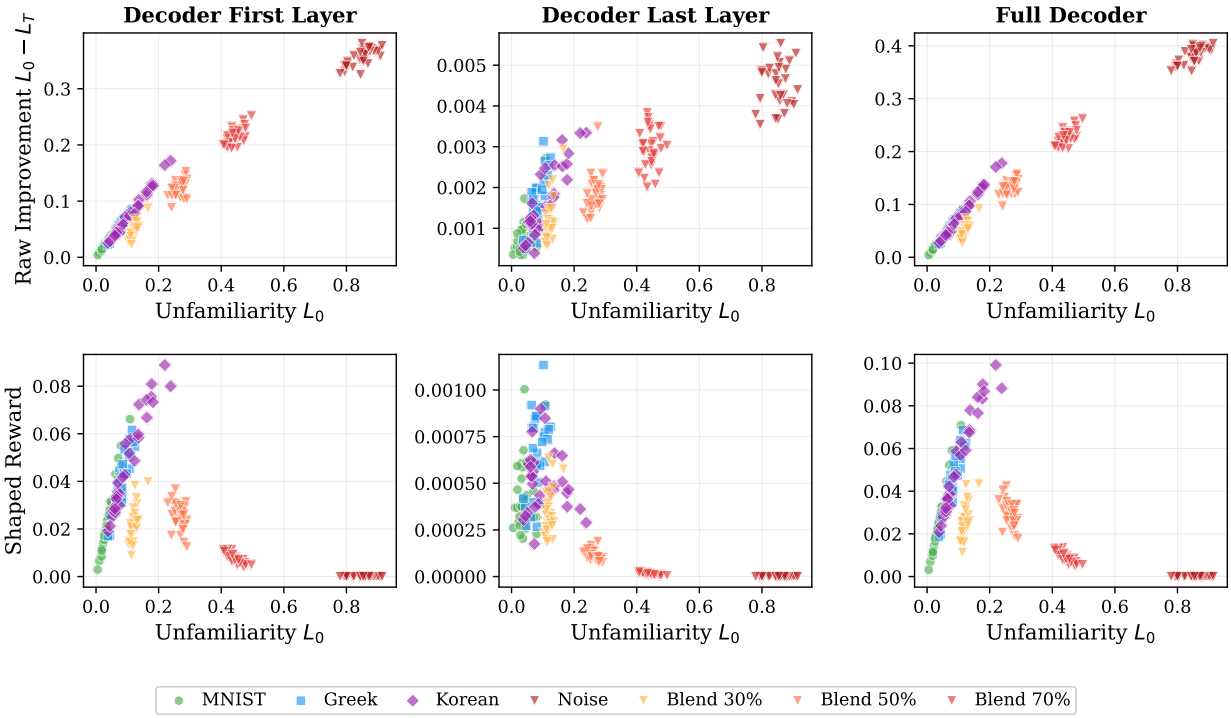


Figure 13: *Layer-specific learnability. The shaped reward when fine-tuning only individual decoder layers during the inner loop, against fine-tuning the full decoder (rightmost). Adapting only the first (high-level, semantic) layer recovers most of the full-decoder reward across categories, while adapting only the last (low-level, textural) layer yields near-zero improvement on every category — learnability is driven by semantic-level adaptation, not by low-level pattern fitting.*

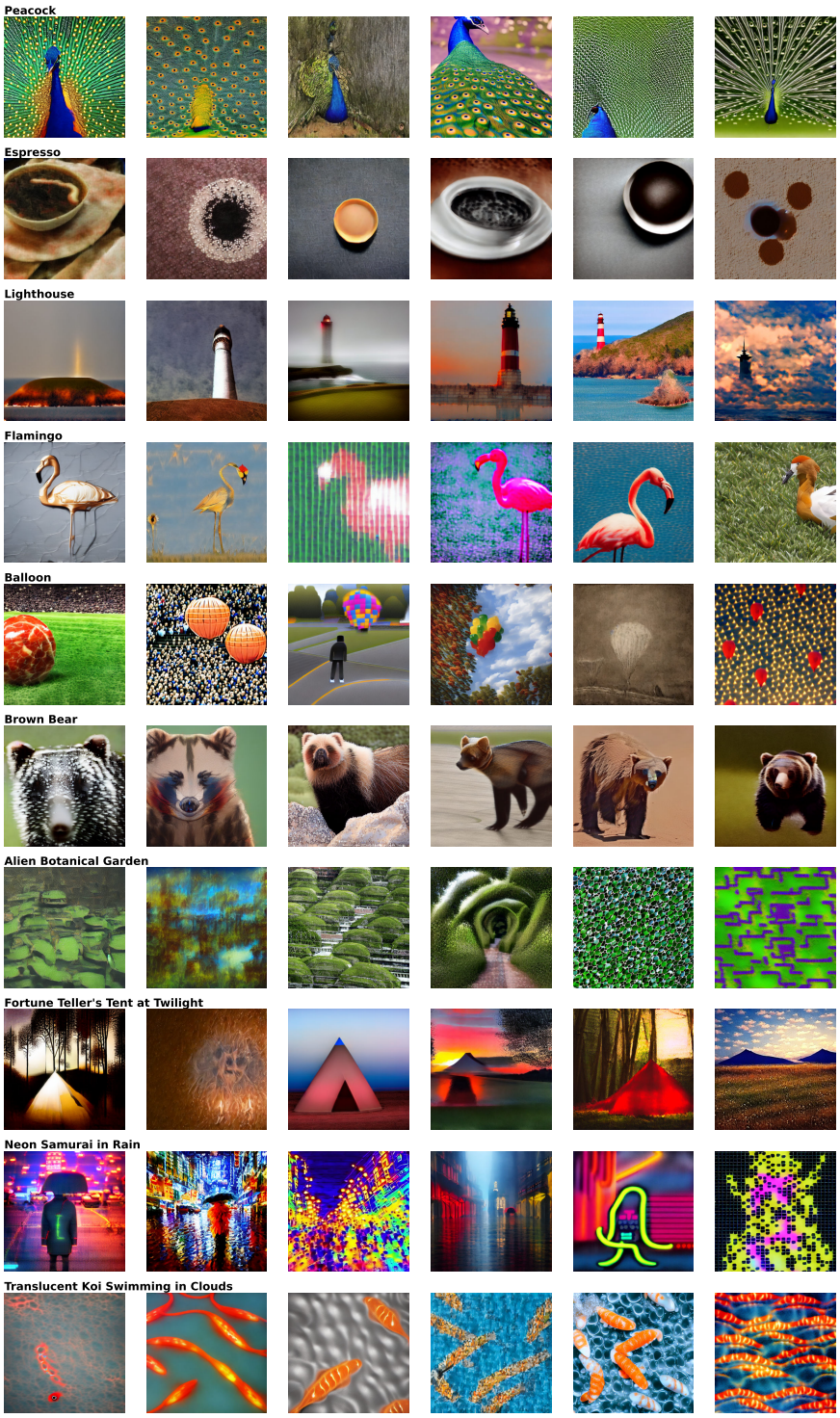


Figure 14: *Automatic top-K outputs from the Creator-Appraiser framework on 10 additional prompts not shown in the main text. Each row is one prompt; six columns are the random sample of top-8 by Δsim within the relevance gate (seed fixed). No manual curation.*

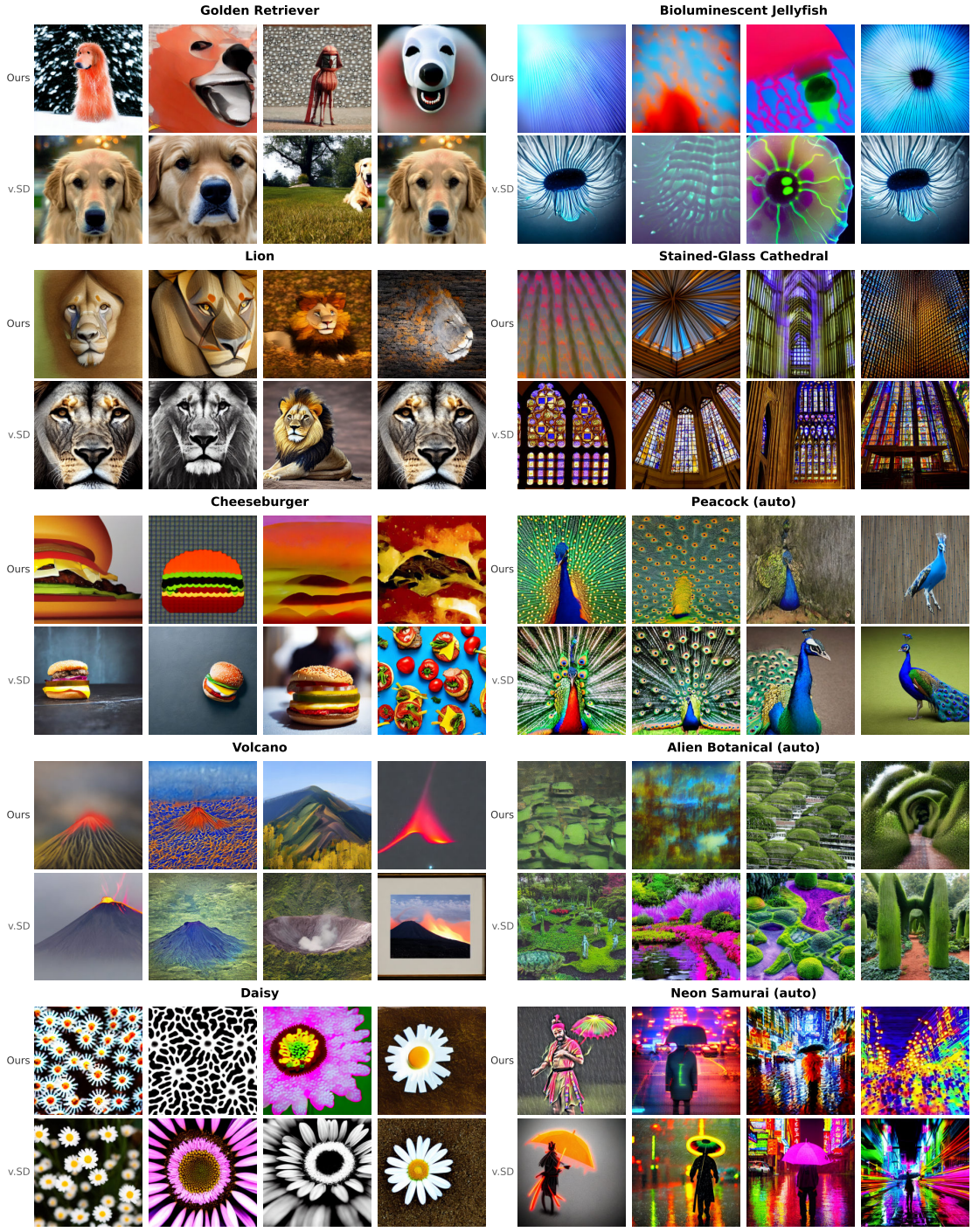


Figure 15: *Each picked Creator-Appraiser sample (top of each pair) shown alongside its DINOv2-nearest neighbor (bottom of each pair) selected from 1000 vanilla Stable Diffusion samples. Top 7 prompts are the manually curated picks from Figure 6; bottom 3 are automatically curated picks from Figure 14 (peacock, an alien botanical garden, neon samurai in rain).*

Table 1: *Per-prompt DINOv2 reachability across LoRA rank r. $\bar{d}_V^V$ is the intra-vanilla nearest-neighbor cosine distance ($|V|$=1000 samples per prompt); the four rank columns give $\bar{d}_V^C$, the mean nearest-neighbor distance from each Creator-Appraiser sample to its nearest vanilla sample, computed over the subset of Creator-Appraiser samples that pass the Qwen2.5-VL relevance score (≥ 5 out of 10).*

	$\bar{d}_V^V$	$\bar{d}_V^C$ at LoRA rank			
Prompt	(intra-V)	r=4	r=8	r=16	r=32
ImageNet class prompts					
balloon	0.199	0.379	0.521	0.518	0.597
brown bear	0.036	0.207	0.477	0.609	0.787
castle	0.285	0.566	0.655	0.708	0.820
cheeseburger	0.084	0.212	0.387	0.419	0.509
daisy	0.160	0.469	0.659	0.740	0.665
espresso	0.102	0.422	0.564	0.662	0.703
flamingo	0.038	0.099	0.206	0.309	0.505
golden retriever	0.104	0.420	0.537	0.675	0.821
jellyfish	0.067	0.388	0.600	0.688	0.709
lighthouse	0.144	0.300	0.341	0.451	0.493
lion	0.030	0.165	0.405	0.399	0.522
peacock	0.014	0.111	0.279	0.440	0.605
volcano	0.068	0.287	0.445	0.584	0.698
Open-ended prompts					
alien botanical garden	0.220	0.631	0.769	0.765	0.796
ancient library lit by fireflies	0.204	0.657	0.736	0.777	0.799
bioluminescent jellyfish	0.122	0.590	0.643	0.664	0.690
fortune teller's tent at twilight	0.189	0.511	0.635	0.756	0.743
neon samurai in rain	0.281	0.488	0.607	0.696	0.729
stained-glass cathedral interior	0.122	0.514	0.679	0.784	0.860
translucent koi swimming in clouds	0.053	0.480	0.627	0.746	0.674
Mean	**0.126**	**0.395**	**0.539**	**0.620**	**0.686**
Median	**0.113**	**0.421**	**0.582**	**0.670**	**0.700**

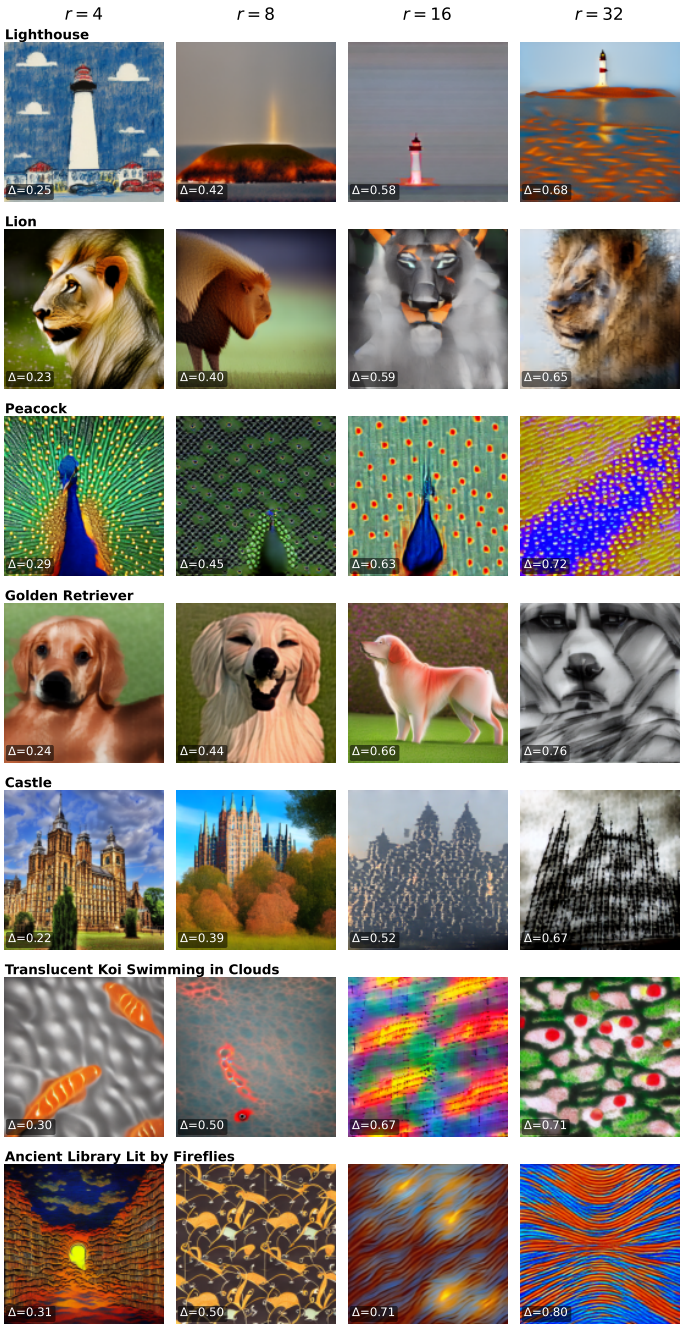


Figure 16: *Effect of LoRA rank: Per prompt, the top Best-of-N survivor (Qwen2.5-VL relevance ≥ 5, ranked by Δsim) at LoRA rank $r \in \{4, 8, 16, 32\}$. As rank grows the inner LoRA gains capacity to assimilate increasingly out-of-distribution outputs, so the meta-gradient is free to push the diffusion sample further from the prompt.*

Table 2: *Per-prompt natural image results, restricted to CA samples passing the Qwen2.5-VL relevance gate (≥ 5, same filter as Table 1). Cells in the rank columns are sim_0 / Δsim, where sim_0 is frozen-CLIP image-text cosine at t=0 (concept fidelity) and Δsim is the inner-loop reward.*

	Vanilla SD	Ours (sim_0 / Δsim)			
Prompt	sim_0	r=4	r=8	r=16	r=32
ImageNet class prompts					
balloon	0.26	0.23 / 0.22	0.20 / 0.38	0.21 / 0.54	0.21 / 0.64
brown bear	0.27	0.21 / 0.23	0.18 / 0.41	0.17 / 0.59	0.11 / 0.73
castle	0.23	0.20 / 0.21	0.20 / 0.36	0.19 / 0.53	0.18 / 0.66
cheeseburger	0.25	0.23 / 0.23	0.22 / 0.40	0.20 / 0.58	0.18 / 0.69
daisy	0.27	0.22 / 0.23	0.20 / 0.41	0.17 / 0.60	0.20 / 0.64
espresso	0.24	0.21 / 0.22	0.21 / 0.38	0.18 / 0.55	0.16 / 0.69
flamingo	0.29	0.25 / 0.24	0.24 / 0.41	0.23 / 0.58	0.18 / 0.68
golden retriever	0.28	0.18 / 0.26	0.18 / 0.44	0.16 / 0.63	0.15 / 0.74
jellyfish	0.27	0.19 / 0.24	0.16 / 0.41	0.14 / 0.61	0.14 / 0.70
lighthouse	0.25	0.22 / 0.22	0.21 / 0.38	0.20 / 0.55	0.20 / 0.66
lion	0.26	0.23 / 0.22	0.22 / 0.38	0.21 / 0.56	0.17 / 0.68
peacock	0.29	0.23 / 0.26	0.21 / 0.43	0.19 / 0.60	0.17 / 0.70
volcano	0.24	0.19 / 0.24	0.19 / 0.39	0.19 / 0.57	0.17 / 0.69
Open-ended prompts					
alien botanical garden	0.27	0.14 / 0.23	0.13 / 0.40	0.12 / 0.58	0.12 / 0.70
ancient library lit by fireflies	0.28	0.13 / 0.28	0.11 / 0.47	0.09 / 0.66	0.08 / 0.75
bioluminescent jellyfish	0.28	0.10 / 0.28	0.10 / 0.47	0.09 / 0.66	0.09 / 0.75
fortune teller's tent at twilight	0.25	0.19 / 0.28	0.16 / 0.46	0.14 / 0.64	0.14 / 0.71
neon samurai in rain	0.34	0.15 / 0.29	0.11 / 0.47	0.08 / 0.66	0.06 / 0.76
stained-glass cathedral interior	0.25	0.17 / 0.26	0.14 / 0.44	0.11 / 0.64	0.08 / 0.76
translucent koi swimming in clouds	0.27	0.18 / 0.29	0.16 / 0.49	0.10 / 0.69	0.19 / 0.67
Mean	**0.27**	**0.19 / 0.25**	**0.18 / 0.42**	**0.16 / 0.60**	**0.15 / 0.70**
Median	**0.27**	**0.20 / 0.24**	**0.19 / 0.41**	**0.17 / 0.59**	**0.16 / 0.70**