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

Mengye Ren
Agentic Learning AI Lab
New York University
mengye@nyu.edu
https://agenticlearning.ai/conceptual-creativity-meta-learning/

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. With a frozen diffusion model and no additional language conditioning, the meta-learning gradient is enough to produce both stylistic variations and concept compositions beyond what the base model can generate on its own.

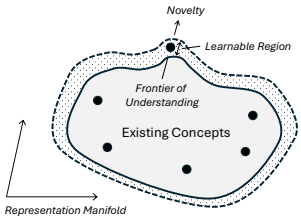
1 Introduction

Humans recognize new concepts from one or a few examples. A single Monet painting is enough to begin grasping Impressionism and to recognize its style in other works. This capacity for few-shot concept learning is a cornerstone of intelligent behavior [9, 27], and its algorithmic counterparts have advanced rapidly through meta-learning [11], prototypical representations [40], and in-context learning [6]. The reverse question has received far less attention: *what does it mean to create a new concept in the first place?*

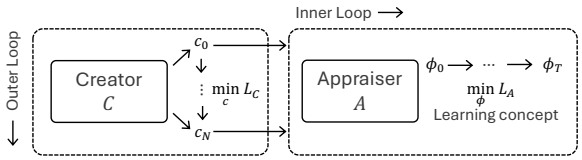
Not every new stimulus is a new concept. An infinite stream of white-noise images is distinct sample by sample, yet each is perceptually indistinguishable from the next and offers no structure to grasp. At the other extreme, a line, a square, or the sequence $1, 2, 3, \ldots$ is fully comprehensible but carries no novelty. A concept worth creating sits between these extremes. This matches the standard definition in the creativity literature, which asks that a creative product be both novel and valuable [5, 35, 38]. For a new concept to be valuable, at least from a perceptual perspective, it must be grounded in our existing representations of the world; it must, on some level, be understandable from past experience.

Whether a stimulus can be assimilated depends on the observer. Many artistic forms, such as Islamic geometric patterns or the repeated motifs of a musical composition, are highly compressible, in the sense that a small fragment suffices to predict the whole. Compressibility, however, is relative to what the observer already knows. Jazz can sound chaotic on first listen and rich once the listener has internalized its chord progressions and rhythmic motifs. Schmidhuber [39] formalized this as *compression progress*: a stimulus is interesting when it improves an observer's predictive model. We adopt the same view of creativity. Existing concepts form a representation manifold built from past experience, and novel concepts lie just beyond it, past what the observer can currently predict but close enough that a small update assimilates them (Figure 1a).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(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*.

We hypothesize that *an artifact is creative for an observer if it is initially unfamiliar to that observer and can be rapidly learned by it through a brief adaptation*. We operationalize unfamiliarity as the observer's initial prediction error L_0, and learnability as the drop from L_0 to the residual error L_T after T adaptation steps. The definition is observer-relative by construction, since the same artifact can be creative for one observer and familiar or unlearnable for another. We use adaptation, learning, and memorization interchangeably, because our focus is the fast adaptation of an observer to a single instance. The empirical claim we test is that the learning progress $L_0 - L_T$ carries information about structured novelty that cannot be recovered from unfamiliarity L_0 alone.

We turn this hypothesis into a generation method through meta-learning. A *Creator* proposes a candidate, and an *Appraiser*, the adaptive observer, takes a few gradient steps to fit it. The Appraiser's improvement is the reward, and the Creator follows the gradient of this reward through the inner loop (Figure 1b). We instantiate the Creator as a frozen diffusion model and the Appraiser as an autoencoder on MNIST and as CLIP with a low-rank text adapter on natural images.

Our contributions are as follows:

- We define creativity as initial unfamiliarity followed by rapid learnability for an adaptive observer, which grounds novelty and value in prediction error.
- We introduce Creator-Appraiser meta-learning, in which a Creator generates artifacts by maximizing an Appraiser's learning progress, and instantiate it on MNIST and on natural images with Stable Diffusion.
- We show that learning progress separates structured novelty from noise on MNIST, and that on Stable Diffusion it leads to interpretations of a prompt that lie outside the model's usual samples.

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 [1, 3, 7, 19, 20, 25, 37, 43].

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 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.

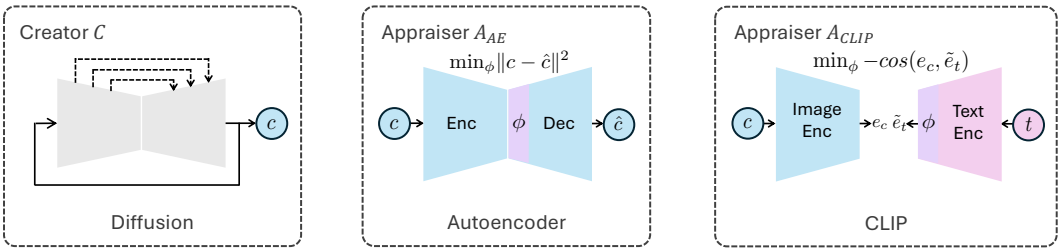


Figure 2: Implementation of the Creator-Appraiser setup in the experiments. We illustrate both the autoencoder and CLIP appraisers. ϕ denotes the adaptation parameter in the Appraiser for concept learning.

Table 1: Appraiser instantiations.

	Autoencoder Appraiser (MNIST)	CLIP Appraiser (natural images)
Creator C	DDPM, frozen	Stable Diffusion v1.5, frozen
Creator variable ξ	noise latent	prompt text embedding
Slow weights θ	LeNet encoder, decoder	CLIP image and text encoders
Fast weights ϕ	first decoder layer	rank-r LoRA on text projection
Inner loss $\mathcal{L}_A$	pixel MSE	negative image-text cosine
Reward R	$(L_0 - L_T)$ with gates g_0, g_T	$\text{sim}_T - \text{sim}_0$
Constraint on R	reward shaping (μ, σ, τ)	adapter rank r
Evaluation	(L_0, L_T) plane, reward ablation	DINOv2 reachability, Qwen2.5-VL relevance

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. The reward to C is thus A's learning progress, and a candidate earns a high reward when A finds it unfamiliar at first and then learns it quickly. Many different candidates can earn a high reward, so the objective does not single out one artifact, and in our experiments different noise seeds lead the Creator to different outputs (Section 3.2). Before each C iteration, we reset the fast weights ϕ to a fixed initialization ϕ_0, so A meets every candidate as a first impression. Algorithm 1 summarizes the procedure, and Figure 1b illustrates it.

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 accepts any adaptive predictive model as the Appraiser. We use an Autoencoder Appraiser for the MNIST experiments and a CLIP Appraiser for the natural image experiments (Figure 2 and Table 1).

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 (guidance scale) η.

$\tilde{c} \leftarrow C(\xi)$ ▷ generate (intermediate or final) candidate
$\phi \leftarrow \phi_0$ ▷ reset Appraiser fast weights for first impression
$L_0 \leftarrow \mathcal{L}_A(\phi, \theta; \tilde{c})$
for $k = 1, \ldots, T$ **do** ▷ inner-loop adaptation, differentiable in ξ
    $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}_A(\phi, \theta; \tilde{c})$
end for
$L_T \leftarrow \mathcal{L}_A(\phi, \theta; \tilde{c})$
$R \leftarrow L_0 - L_T$ ▷ 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.), and 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 T 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 a Gaussian complexity gate and an exponential success gate, which constrain $\mathcal{L}_{A,0}$ and $\mathcal{L}_{A,T}$ to a "learnable yet unfamiliar" region:

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

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

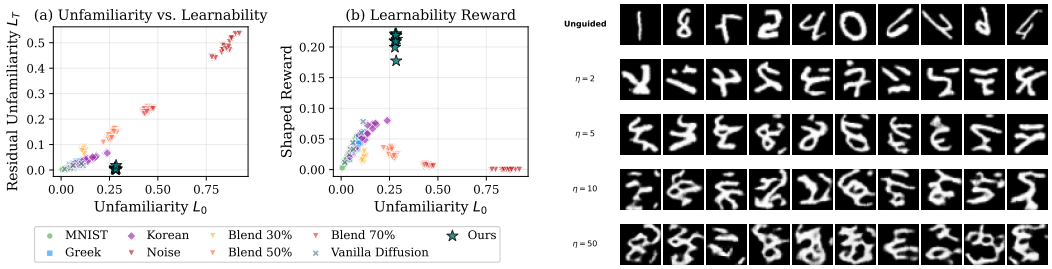
The *complexity gate* g_0 keeps the initial prediction error in a target range around μ (avoiding both familiar low-$\mathcal{L}_{A,0}$ inputs and noise-like high-$\mathcal{L}_{A,0}$ inputs). The *success gate* g_T rewards low final error, ensuring the decoder actually fits the input. We show the effect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

The CLIP Appraiser measures how quickly a small text-side update can learn to match a generated image (Figure 2). We use the contrastive image-text model CLIP [33] with both encoders frozen. The inner loop adapts the text projection layer with a low-rank adapter [22] of rank r, $W'_{\text{text}} = W_{\text{text}} + UV$, with U initialized to zero. Over T steps, it updates U and V to maximize $\cos(e_{\text{img}}, \tilde{e}_{\text{text}})$, where $\tilde{e}_{\text{text}}$ is the adapted text feature. The reward is the gain in image-text similarity, $R = \text{sim}_T - \text{sim}_0$.

The rank r limits the capacity of the adaptation and takes the place of the Gaussian gates of the Autoencoder Appraiser. At small r, the inner loop cannot fit an arbitrary image direction in T steps, so the meta-gradient stays near the prompt's concept. A larger r relaxes this limit and lets the outputs



(a) Unfamiliarity–learnability space.

(b) Guided samples at increasing η.

Figure 3: **(a)** Guided images are moderately unfamiliar yet quickly learnable, and their shaped reward exceeds that of the best reference category (Korean). **(b)** Larger η gives more novel and complex samples that keep a stroke-like structure. At the largest η, they no longer match any digit.

move farther from the prompt (Section 3.2). With this limit in place, we use the similarity gain as the reward without further shaping.

Dependence on the Appraiser. Because creativity is defined relative to an observer, the Appraiser's prior and loss decide which artifacts count as creative (Table 1). The framework accepts any predictive loss, but each loss induces its own notion of creativity. A reconstruction loss rewards strokes and textures the decoder can reproduce, and a classifier loss rewards movement across fixed class boundaries. The CLIP loss rewards images whose relation to the prompt a small text-side update can capture, which favors re-interpretations of the concept. In preliminary experiments, DINOv2-based reconstruction and classifier Appraisers produced degenerate images, likely because the reconstruction objective is low-level and the classifier lacks constraints and neighboring concepts to contrast with.

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 [21] 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. In Figure 3a, we probe whether the shaped reward correctly orders stimuli that span the unfamiliarity–learnability space, comparing 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 blends have similar L_0 to the Omniglot characters but lack coherent structure. In the (L_0, L_T) plane (left panel), noise lies in the top-right corner, unfamiliar and unlearnable. Familiar digits lie in the bottom-left, and our guided generations lie in the bottom-middle, moderately unfamiliar yet quickly learnable. The shaped reward (right panel) ranks Korean and Greek above noise blends with the same L_0 and gives noise a near-zero reward.

Hypothesis test. This experiment tests the central claim of the paper, that the improvement term $L_0 - L_T$ adds signal beyond unfamiliarity L_0. The null hypothesis is that the shaped reward reduces to gated unfamiliarity. Figure 4 ablates each reward component. In Figure 4a, the complexity gate alone cannot separate Korean characters from noise-blended digits at the same L_0, because both pass the gate. In Figure 4b, adding the improvement term separates them, since the decoder reduces its loss on Korean characters much more than on noise blends with comparable initial error. In the

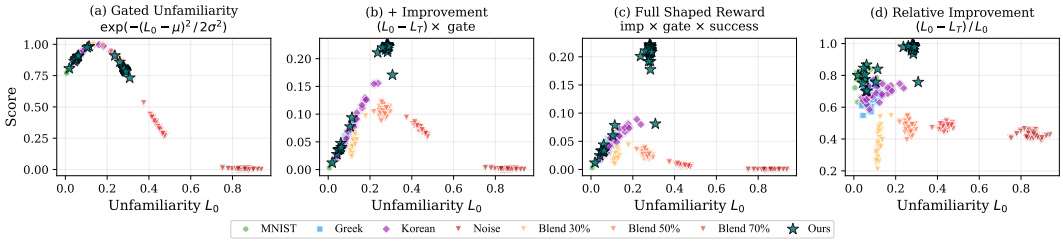


Figure 4: Decomposition of the shaped reward. **(a)** The complexity gate alone cannot distinguish structured from degraded stimuli at similar L_0. **(b)** Adding the improvement term separates structured scripts from noise blends. **(c)** The full shaped reward. **(d)** Relative improvement inverts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

overlap zone $L_0 \in [0.05, 0.15]$, the rank correlation with the shaped reward is $\rho = -0.10$ for the gate and $\rho = 0.89$ for the improvement term, which contradicts the null hypothesis. In Figure 4c, the full reward adds the success gate g_T to suppress inputs the decoder cannot fit. In Figure 4d, the *relative* improvement $(L_0 - L_T)/L_0$ reverses the ranking and rewards familiar inputs over structured novel ones.

Layer-specific learnability. We further ablate which decoder layers are fine-tuned during the inner loop (Appendix Figure 14). 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, so learnability is driven by adaptation of semantic features.

Guidance scale and ideal initial loss. We vary the guidance scale η and the complexity-gate target μ (Appendix Figures 10 and 12). A larger η gives more learnable but visually busier samples. Raising μ increases L_0 while L_T stays nearly constant, so the gate sets the level of unfamiliarity without imposing a hard threshold.

Comparison with baselines. We compare against three baselines. **Vanilla DDPM** samples without guidance. **Noise perturbation** adds noise of scale $\sigma_p \in \{2, 5, 10, 20\}$ to the latent at each step. **Repulsive guidance** [29] pushes the predicted latent away from class prototypes with scale $\eta \in \{1, 2, 5, 10, 20, 50, 100\}$, which gives deviation without learnability. In Figure 5a, noise perturbation stays near the origin because the DDPM absorbs the perturbation. Repulsive guidance follows the no-learning diagonal $L_T \approx L_0$, and only our method reaches the bottom-right. In Figure 5b, noise-perturbed samples look like unguided digits, repulsive-guided samples are disrupted but unstructured, and ours are coherent stroke-like glyphs. Additional samples from each method are in Appendix Figure 9.

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 [36]. 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. We found this search in semantic space necessary for the reward to track concept-level fit rather than pixel texture. We evaluate two types of 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 two quantities from the CLIP Appraiser and two measures from separate evaluator models. The Appraiser quantities are (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, which is also the inner-loop initial similarity, and (ii) $\Delta\mathbf{sim} = \mathbf{sim}_T - \mathbf{sim}_0$, the similarity improvement after the T-step LoRA inner loop, which is the reward the meta-gradient optimizes. The evaluators do not take part in generation. (iii) The **relevance rate** is the fraction of samples that Qwen2.5-VL-7B [2], a

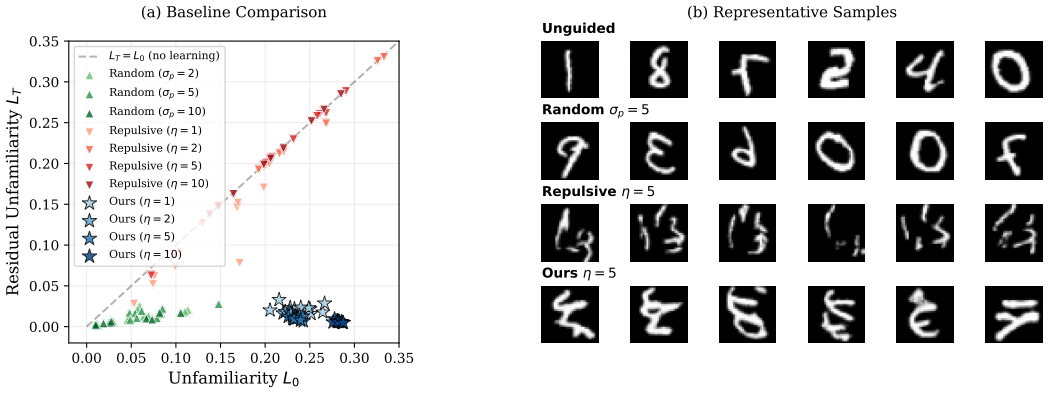


Figure 5: Baseline comparison. **(a)** Each point is one generated image in the (L_0, L_T) plane; the dashed diagonal marks $L_T = L_0$ (no learning). Color gradients show increasing η (or σ_p) within each method. In the legend, Random is noise perturbation and Unguided is vanilla DDPM. Noise perturbation 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 noise perturbation).

vision-language model (VLM), scores at least 5 on a 0–10 scale of concept presence (Appendix B). (iv) The **DINOv2 distance** is the nearest-neighbor distance from a sample to 1000 vanilla samples of the same prompt in DINOv2 feature space [31], described below.

Results. Figure 6 shows representative picks; the vanilla SD baselines on the right are uniformly photorealistic and on-prompt. The picks include forms that the prompt does not describe: a lion in autumn-leaf form, golden retrievers in novel coloration and proportion, a cheeseburger rendered as a sunset landscape, a daisy whose petals dissolve into ink strokes, and a pastel oil-painted volcano. 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 deviation is found inside the prompt embedding by the meta-gradient on the inner-loop learning-progress signal.

Capacity-vs-fidelity trade-off. As rank grows (Figure 7), Δsim rises while sim_0 and the relevance rate fall. A low-rank inner LoRA cannot fit off-manifold images, which keeps the meta-gradient near the prompt's class; higher rank lifts that constraint.

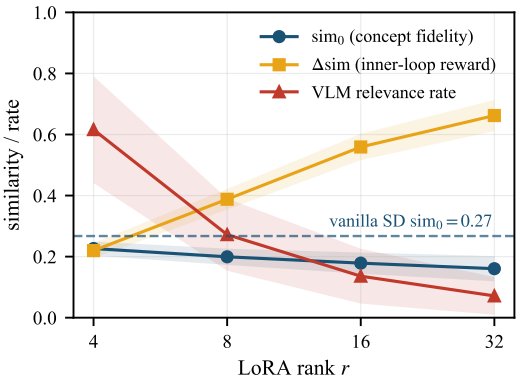


Figure 7: Effect of LoRA rank. Lines show the mean ± 1 s.d. across 20 prompts. As r grows, the inner-loop reward Δsim rises, frozen-CLIP sim_0 drifts below the vanilla level, and the Qwen2.5-VL relevance rate falls. Per-prompt values in Table 6.

Reachability beyond vanilla. We ask whether our outputs are samples that vanilla Stable Diffusion would also produce given enough draws. For each of the 20 prompts we draw $N = 1000$ vanilla samples, embed them and our samples with a frozen DINOv2 ViT-L/14 [31], and compute nearest-neighbor cosine distance to the vanilla set. We keep only our samples with Qwen2.5-VL relevance ≥ 5. At r=8, our samples lie a median $4.6\times$ farther from the vanilla set than vanilla samples lie from each other, and this distance grows with LoRA rank. The meta-gradient therefore reaches outputs that vanilla sampling does not reach at $N = 1000$. Table 2 compares baselines on the same prompts, with samples in Figure 8. In those samples, CFG s=1 outputs are desaturated and blurry, ProCreate γ=500 outputs often become flat graphic shapes, and ours vary in color, texture, and composition. Baselines that lower the classifier-free guidance (CFG) scale or repel samples from reference images reach a similar trade-off between distance and relevance rate. Our method reaches

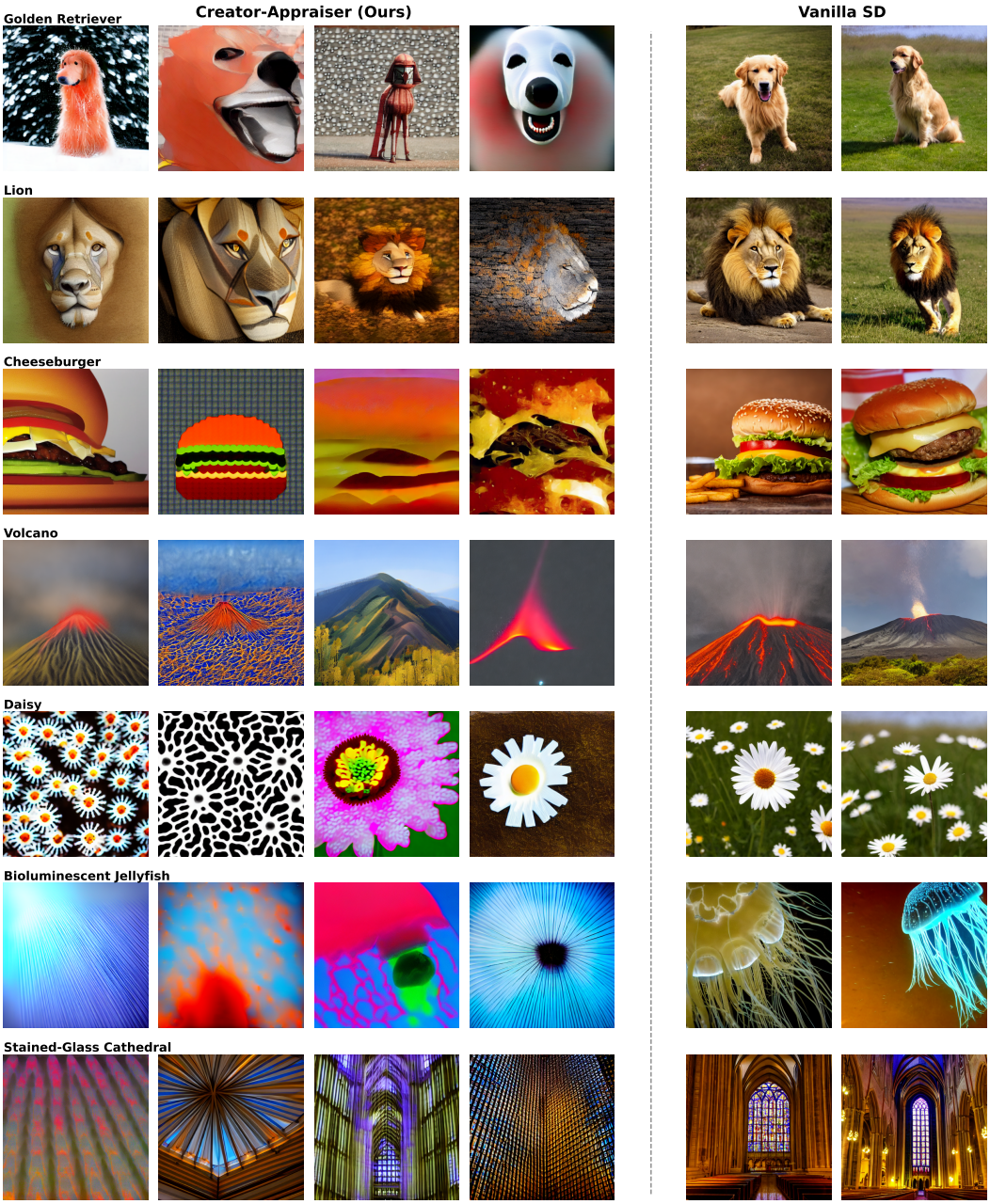


Figure 6: Curated outputs from the CLIP Appraiser at LoRA rank $r \in \{8, 32\}$, $\eta = 10$. Each row shows four samples from our method (left) and two vanilla SD samples (right) for the same prompt; rows 1–5 are ImageNet class-conditioned, rows 6–7 are open-ended text prompts. The meta-gradient produces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

it through the learnability objective alone, without an explicit deviation term. The increase with rank also holds without the relevance gate and at a fixed sample count per prompt, and Appendix C gives per-prompt numbers and nearest-neighbor pairs. Appendix B shows uncurated top-K samples.

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"

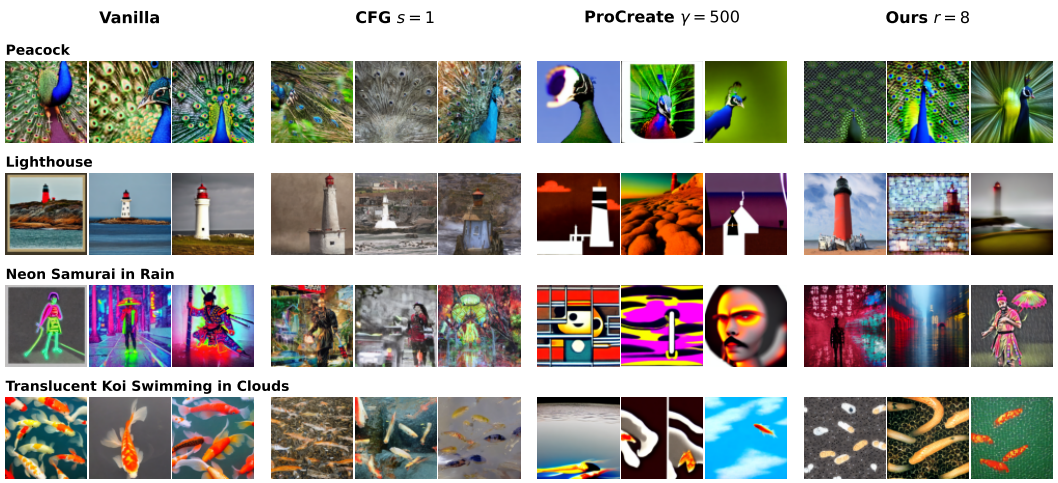


Figure 8: Baselines and ours (r=8) on four prompts. Samples are random draws with Qwen2.5-VL relevance ≥ 5 (vanilla ungated).

Table 2: Baselines on 20 prompts. Rate: fraction with Qwen2.5-VL relevance ≥ 5. NN dist.: median DINOv2 distance of those samples to the 1000 vanilla samples. Vanilla distance is to the other 999 samples. Full results in Appendix Table 4.

Baseline	Rate	NN dist.	Ours	Rate	NN dist.
Vanilla (CFG 7.5)	0.96	0.11	$r = 4$	0.62	0.41
CFG $s \in \{3, 5, 10\}$	0.94–0.97	0.06–0.17	$r = 8$	0.27	0.53
CFG $s = 1$	0.64	0.47	$r = 16$	0.14	0.55
CFG $s = 0.5$	0.17	0.66	$r = 32$	0.07	0.66
ProCreate [29], $\gamma = 250$	0.73	0.38			
ProCreate, $\gamma = 500$	0.28	0.58			
ProCreate, $\gamma = 1000$	0.08	0.72			

[28]. We take the position that creativity is best defined operationally, by its *effect on an observer* rather than by the creator's experiential phenomenology [30]. 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 [13]). Boden [5] distinguishes *exploratory* creativity, which works within an existing conceptual space, from *transformational* creativity, which reshapes the space itself. We do not assign outputs to one of these categories. In our experiments the Creator is frozen, so its outputs explore an existing space, and the Appraiser learns each output in its inner loop.

Schmidhuber [39] 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 [41]) aimed at RL task completion rather than visual creation. Our reward shaping also resonates with Berlyne [4]'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. *Bayesian surprise* [23] measures how much an observation moves an observer, as the KL divergence between posterior and prior beliefs under a fixed probabilistic model. Our reward instead measures the loss reduction from a bounded gradient update, and the Appraiser's loss plays the role of the likelihood. Classical information measures abstract away the bounded observer: Shannon entropy treats pure noise as maximally informative, while Kolmogorov complexity [26] assumes an unbounded observer for whom any compressible structure is free. Neither captures structure a *bounded* learner can extract with a small adaptation. Most closely

related, Finzi et al. [12] 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, whereas 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. GANs [16, 24] and diffusion models [21] are the dominant paradigm for high-fidelity image generation. Creative Adversarial Networks [8] extended GANs by training the generator to deviate from learned artistic styles while remaining aesthetically plausible. However, the apparent creativity of these models is, by default, largely a consequence of recombining concepts and styles already present in the training distribution [10, 34], 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. [29] guide diffusion away from existing training examples using a propulsive energy term that repels generated samples from specific reference images. Han et al. [18] enhance creativity by amplifying low-frequency components in early diffusion layers. Golan et al. [15] use vision-language models to adaptively construct negative prompts that steer generation toward unconventional outputs. These methods share a strategy of pushing the generator away from its typical outputs. We instead optimize for *learnability*: the output must be unfamiliar to the observer but also structured enough to be learned quickly.

Adversarial, actor-critic, and meta-learning frameworks. The Creator-Appraiser structure resembles the generator-discriminator dynamic in GANs [16] and the actor-critic framework in RL [42], with 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 [32, 39]. The inner-loop adaptation draws on meta-learning [11], and the text-embedding optimization relates to Textual Inversion [14], though with a learnability rather than reconstruction objective.

5 Conclusion

We started from the view that an artifact is creative for an observer when it lies just beyond what the observer can predict but remains learnable. Our Creator-Appraiser framework turns this view into a meta-learning objective, since the Creator can follow the Appraiser's learning progress through the inner loop. We tested the framework with MNIST digits and natural images. On MNIST, learning progress separates structured from unstructured stimuli, and in both settings its meta-gradient moves a frozen diffusion model away from its typical outputs. As future work, an Appraiser with memory across creations could judge an artifact learnable in light of what it has recently seen, which opens room for context-dependent and curriculum-driven creativity. The Creator-Appraiser structure also applies to any generator paired with an adaptive predictor, and extending to more modalities is the natural next step.

Limitations. Our observers are computational: an autoencoder and a CLIP Appraiser, with DINOv2 and Qwen2.5-VL as evaluators. We have not tested whether people judge the outputs as creative. The CLIP Appraiser also does not enforce fidelity to the prompt, and the relevance rate falls as the rank grows (Table 2). A relevance term in the reward could keep more samples on-prompt. Finally, our framing covers perceptual value only, rather than functional value.

Acknowledgements and Disclosures

This work was supported in part by Visko Platform, Pionex, a Google TPU Award, and the Institute of Information & Communications Technology Planning Evaluation (IITP) under grants RS-2024-00469482 and RS-2024-00509279, 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.

Funding Transparency Statement

Funding in direct support of this work: Visko Platform, Pionex, a Google TPU Award, IITP grants RS-2024-00469482 and RS-2024-00509279, and NYU High Performance Computing.

Additional revenues in the 36 months before submission, not used for this work: the NYU-KAIST Award A25-0081-002, student support from the Meta AI Mentorship Program, Azure credits from the Microsoft Accelerating Foundation Models Research program, Google Cloud credits, Anthropic API credits, and compute from the Center for AI Safety.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 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)*, 2023.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3(Feb):1137–1155, 2003.

[4] Daniel E. Berlyne. *Conflict, Arousal, and Curiosity*. McGraw-Hill, 1960. doi: 10.1037/11164-000.

[5] Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004. doi: 10.4324/9780203508527.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 1877–1901, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

[8] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[9] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. doi: 10.1109/TPAMI.2006.79.

[10] Fu Feng, Yucheng Xie, Xu Yang, Jing Wang, and Xin Geng. Distribution-conditional generation: From class distribution to creative generation. *arXiv preprint arXiv:2505.03667*, 2025.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*. PMLR, 2017.

[12] Marc Finzi, Shikai Qiu, Yiding Jiang, Pavel Izmailov, J. Zico Kolter, and Andrew Gordon Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv preprint arXiv:2601.03220*, 2026.

[13] Giorgio Franceschelli and Mirco Musolesi. Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41, 2024. doi: 10.1145/3664595.

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

[15] Shelly Golan, Yotam Nitzan, Zongze Wu, and Or Patashnik. Vlm-guided adaptive negative prompting for creative generation. *arXiv preprint arXiv:2510.10715*, 2025.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[17] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022.

[18] Jiyeon Han, Dahee Kwon, Gayoung Lee, Junho Kim, and Jaesik Choi. Enhancing creative generation on stable diffusion-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[20] Geoffrey E Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[23] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49 (10):1295–1306, 2009. doi: 10.1016/j.visres.2008.09.007.

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

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[27] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2011.

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

[29] Jack Lu, Ryan Teehan, and Mengye Ren. Procreate, don't reproduce! propulsive energy diffusion for creative generation. In *Computer Vision - ECCV 2024*. Springer, 2024. doi: 10.1007/978-3-031-73027-6_23. URL `https://doi.org/10.1007/978-3-031-73027-6_23`.

[30] Bence Nanay. An experiential account of creativity. In Elliot Samuel Paul and Scott Barry Kaufman, editors, *The Philosophy of Creativity: New Essays*. Oxford University Press, 2014. doi: 10.1093/acprof:oso/9780199836963.003.0002.

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[32] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, 2017.

[33] 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 *International Conference on Machine Learning (ICML)*, 2021.

[34] Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics*, 43(3):1–14, 2024. doi: 10.1145/3659578.

[35] Graeme Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99, 2007. doi: 10.1007/s11023-007-9066-2.

[36] 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)*, 2022.

[37] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press, 1986. doi: 10.7551/mitpress/4943.003.0128.

[38] Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

[39] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368.

[40] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[41] Rupesh Kumar Srivastava, Bas R. Steunebrink, and Jürgen Schmidhuber. First experiments with PowerPlay. *Neural Networks*, 41:205–212, 2013. doi: 10.1016/j.neunet.2013.01.022.

[42] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

[43] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1096–1103, 2008. doi: 10.1145/1390156.1390294.

Appendix

A Additional MNIST Results

This section gives MNIST samples for each baseline (Figure 9), the effect of the guidance scale η and the gate target μ (Figures 10–13), and the layer ablation (Figure 14).

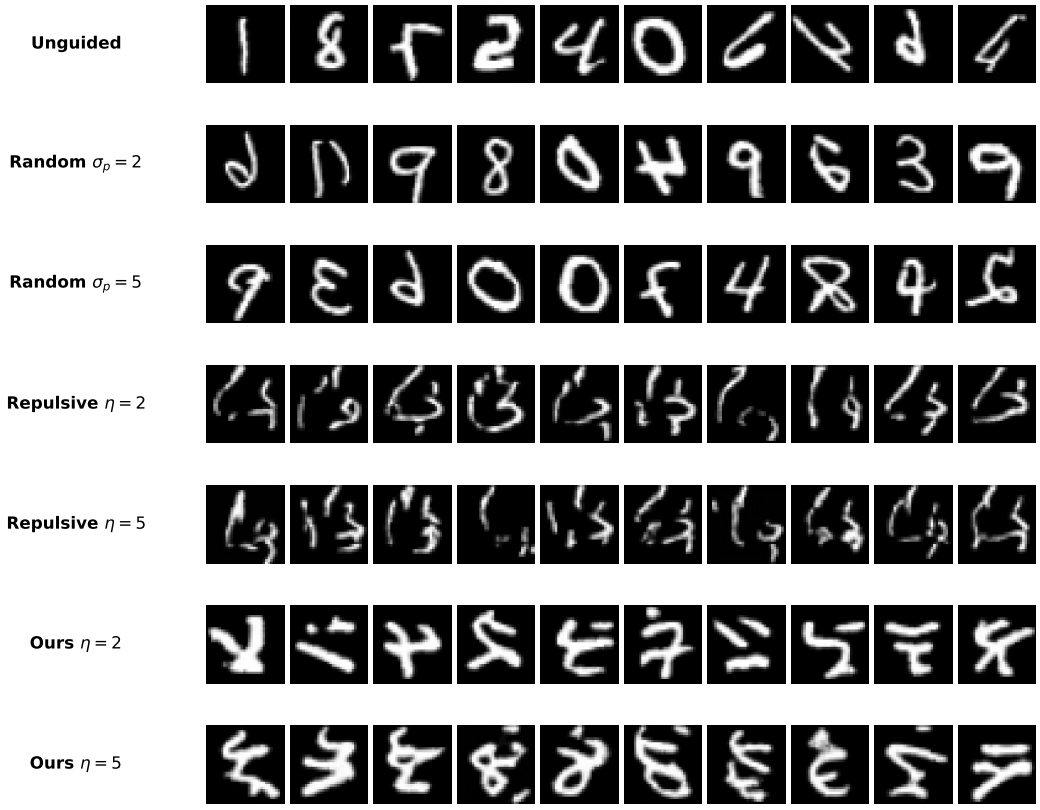


Figure 9: Representative images from each method. **Unguided** (vanilla DDPM): standard DDPM digits. **Random** (noise perturbation, $\sigma_p = 2, 5$): outputs look like unguided samples, because the diffusion process absorbs the noise. **Repulsive guidance** ($\eta = 2, 5, 10$): images become more corrupted, and higher repulsion gives more unfamiliar and less coherent outputs. **Ours** ($\eta = 2, 5$; $\mu = 0.08$): novel, glyph-like patterns with coherent stroke structure.

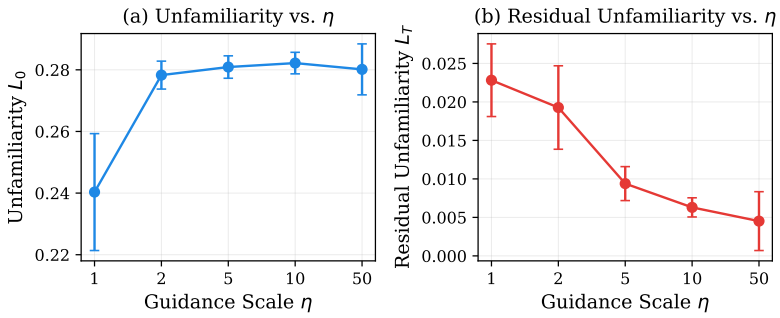


Figure 10: Effect of guidance scale η on generated image statistics ($\mu = 0.15$). **(a)** Unfamiliarity L_0 saturates quickly beyond $\eta = 2$. **(b)** Residual unfamiliarity L_T falls as η grows, so outputs become more learnable.

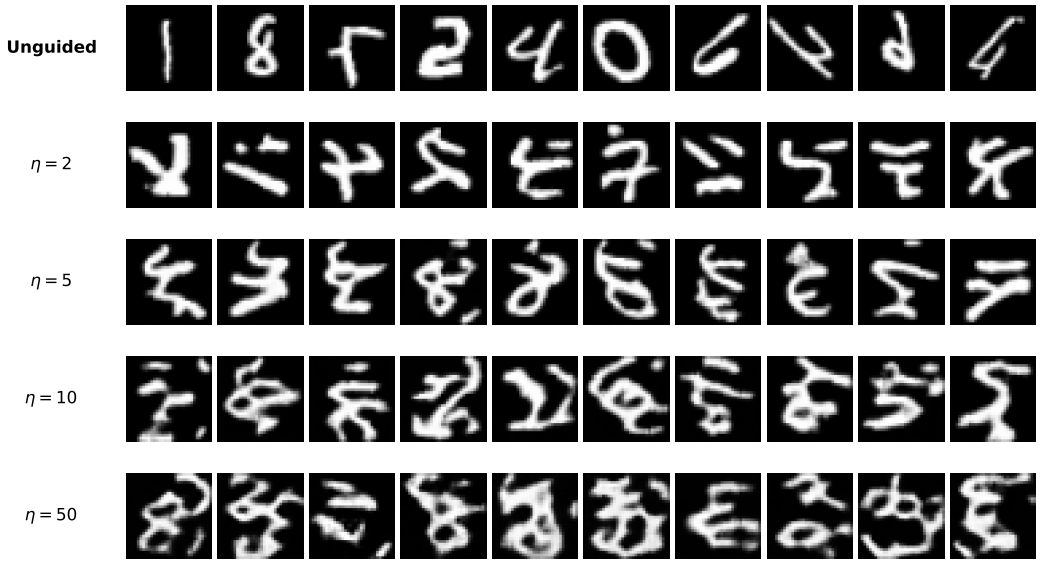


Figure 11: Generated images at increasing guidance scales η, corresponding to the analysis in Figure 10. Vanilla DDPM samples resemble MNIST digits. Guided outputs become more novel and keep a stroke-like structure.

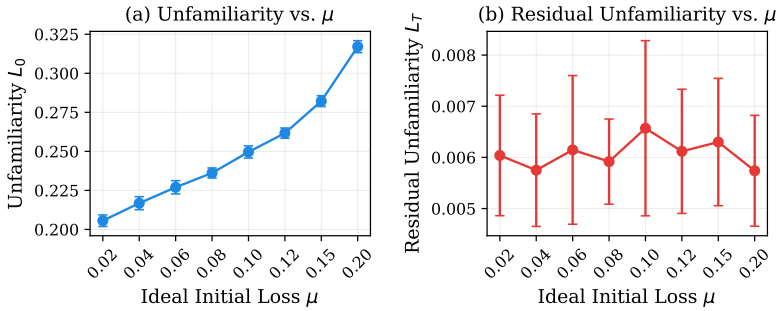


Figure 12: Effect of ideal initial loss μ on generated image statistics ($\eta = 10$). **(a)** Unfamiliarity L_0 grows linearly with μ, so the complexity gate sets the level of unfamiliarity. **(b)** Residual unfamiliarity L_T stays near 0.006 for every μ.

B Automatic Top-K Pipeline and Outputs Across More Prompts

The samples in Figure 6 were picked by hand. Figure 15 shows samples picked by an automatic pipeline on 10 further prompts.

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 100 samples per (prompt, rank) cell at outer-loop guidance $\eta = 10$, $T = 10$ 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 T-step LoRA inner loop on the saved image. This re-scores the image against the same predictor used during generation.

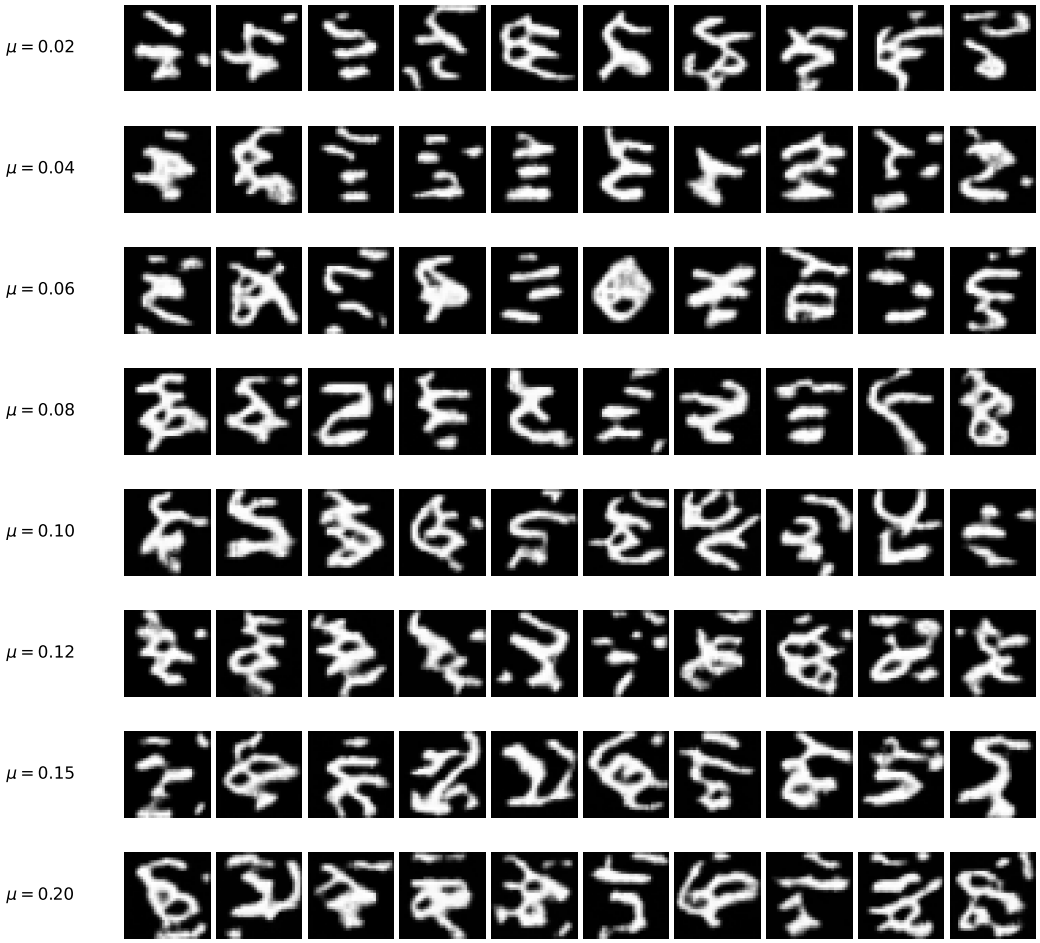


Figure 13: Generated images across ideal initial loss μ values ($\eta = 10$), corresponding to the analysis in Figure 12. Lower μ gives thinner, glyph-like forms, and higher μ gives thicker strokes that depart further from the training digits.

- A Qwen2.5-VL-7B-Instruct relevance score [2] 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 rejects broken outputs, and the VLM gate checks relevance. Survivors are ranked by Δsim in descending order; the top-8 are kept as candidate picks.

Sampling for the figure. For each prompt we show a random 6 of the top 8, with a fixed seed, so the figure is not filled with near-copies of the single highest-Δsim image.

C Reachability: DINOv2 Distance from the Vanilla Samples

We draw $N = 1000$ vanilla samples per prompt, embed both vanilla samples and our samples with a frozen DINOv2 ViT-L/14 [31], L_2-normalize, and report nearest-neighbor cosine distances, mirroring the per-sample rarity construction of Han et al. [17]. For one of our samples x, $d_V(x) = \min_{x' \in V} \left[1 - f(x)^\top f(x')\right]$ measures how far x is from the vanilla set V; for a vanilla sample x', $d_V(x') = \min_{x'' \in V \setminus \{x'\}} \left[1 - f(x')^\top f(x'')\right]$ is the typical distance within V.

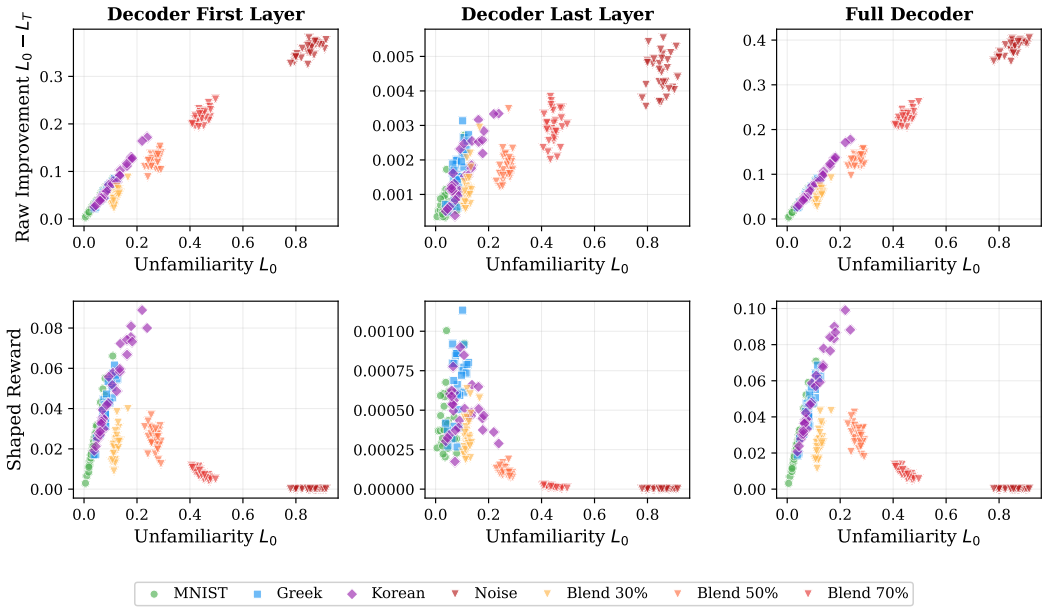


Figure 14: Layer-specific learnability. The shaped reward when fine-tuning only individual decoder layers during the inner loop, against fine-tuning the full decoder (rightmost). Adapting only the first (high-level, semantic) layer recovers most of the full-decoder reward across categories, while adapting only the last (low-level, textural) layer yields near-zero improvement on every category.

Table 5 gives the per-prompt distances for our method and the diversity baselines. At r=8, the ratio of our distance to the intra-vanilla distance ranges from 1.9× (neon samurai in rain) to 17× (peacock).

This reachability metric is not itself a creativity score. DINOv2 distance measures atypicality relative to vanilla sampling, while the Appraiser's inner-loop reward supplies the learnability criterion and the frozen-CLIP sim_0 and Qwen2.5-VL relevance check tie the sample to the prompt. We read the three together as "unfamiliar but memorable".

Selection controls. The relevance gate keeps fewer samples at higher rank, so the gated medians are computed over smaller subsets. Table 3 repeats the analysis without the gate and with a fixed count, keeping the 13 samples per prompt with the highest Qwen2.5-VL relevance score at every rank. The distance increases with rank under all three protocols, and the ungated distances are higher than the gated ones, so the gate makes the reported numbers conservative. The atypical tail of the vanilla pool stays on-prompt, and only the single most atypical of 1000 vanilla samples approaches the distance of our gated samples at r=8.

Table 3: DINOv2 distance to the vanilla pool, median over 20 prompts (18 for gated r=32). Top: selection protocols for our samples. Bottom: vanilla samples by atypicality.

	r=4	r=8	r=16	r=32
Gated (relevance $\geq$ 5)	0.41	0.53	0.55	0.66
Ungated (all 100 samples)	0.57	0.73	0.81	0.84
Top 13 by relevance	0.40	0.53	0.65	0.74
Vanilla	Typical	Top 10%	Top 5%	Most atypical
Distance	0.11	0.17	0.21	0.52
Relevance rate	0.96	0.96	0.95	0.85

Compute cost. On one H200 GPU with 50 denoising steps, our method at r=8 with T=10 inner-loop steps takes 7.7 s per image, against 0.86 s for vanilla Stable Diffusion, about 9× more. We time each method at 2 and 20 images and divide the difference by 18, which removes model loading. The cost is expected to grow linearly with the number of inner-loop and denoising steps.

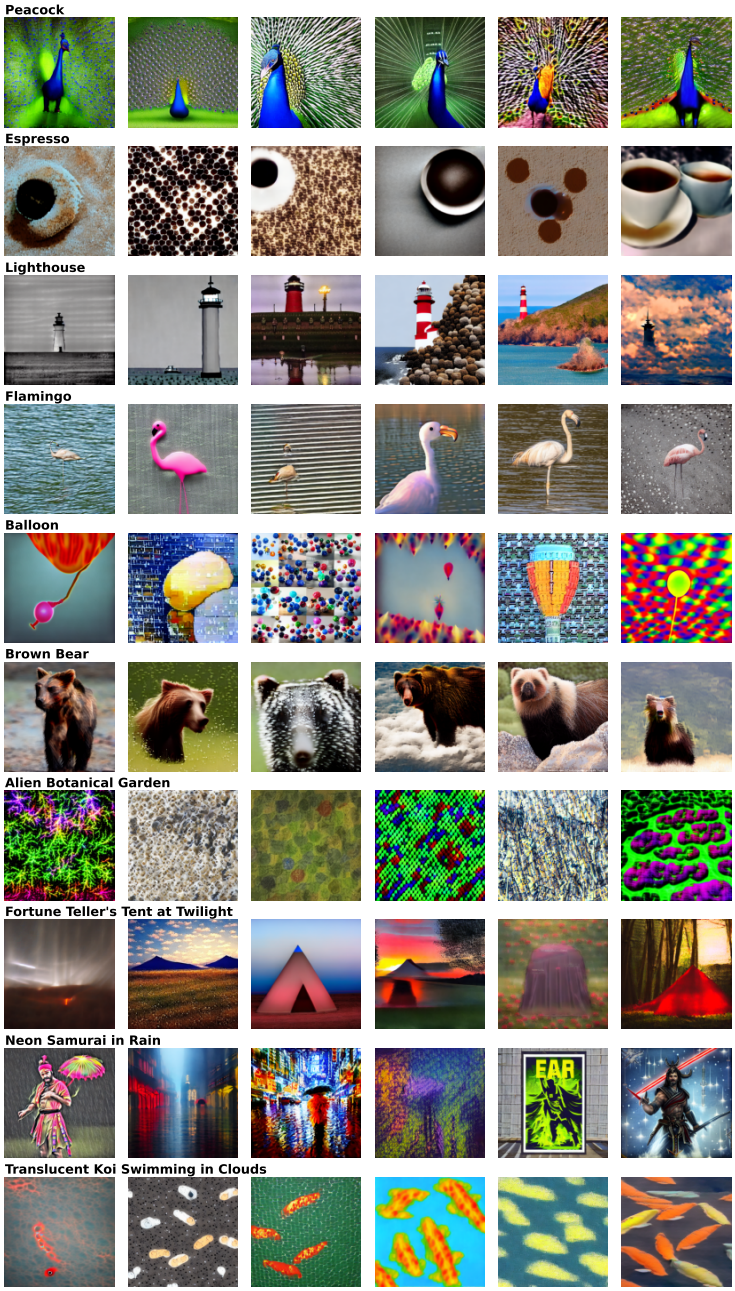


Figure 15: Automatic top-K outputs from the Creator-Appraiser framework on 10 additional prompts not shown in the main text. Each row is one prompt; six columns are the random sample of top-8 by Δsim within the relevance gate (seed fixed). No manual curation.

Baseline samples. Figure 16 shows samples of the diversity baselines next to ours, paired at matched relevance rate. CFG s=1 samples are desaturated and blurry, ProCreate γ=500 samples often become flat graphic shapes, and our samples vary in color, texture, and composition.

Figure 17 pairs each picked sample with its DINOv2-nearest vanilla counterpart from the 1000-sample set. In the rightmost pair of the Daisy row of Figure 17, 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.

Table 4: Baselines on 20 prompts. Relevance rate: fraction with Qwen2.5-VL relevance ≥ 5. NN distance and sim_0: medians over prompts on those samples, with distance to the 1000 vanilla samples (vanilla: to the other 999). ProCreate uses CLIP features in place of the original DreamSim features. Its γ=1000 row and our r=32 row cover 19 and 18 prompts.

Method	Relevance ≥ 5 rate	NN distance	sim_0
Vanilla (CFG 7.5)	0.96	0.11	0.27
CFG scale $s \in \{3, 5, 10\}$	0.94–0.97	0.06–0.17	0.26–0.27
CFG scale $s = 1$	0.64	0.47	0.23
CFG scale $s = 0.5$	0.17	0.66	0.20
Negative prompt (4 variants)	0.95–0.96	0.10–0.12	0.26
ProCreate [29], $\gamma = 250$	0.73	0.38	0.21
ProCreate, $\gamma = 500$	0.28	0.58	0.18
ProCreate, $\gamma = 1000$	0.08	0.72	0.15
Ours, $r = 4$	0.62	0.41	0.23
Ours, $r = 8$	0.27	0.53	0.20
Ours, $r = 16$	0.14	0.55	0.18
Ours, $r = 32$	0.07	0.66	0.16

Table 5: Per-prompt mean DINOv2 nearest-neighbor distance to 1000 vanilla samples, on samples with Qwen2.5-VL relevance ≥ 5. Intra-V: distance of each vanilla sample to the other 999. PC: ProCreate. –: no sample passes.

		Baselines					Ours, LoRA rank			
Prompt	Intra-V	CFG 1	CFG 0.5	PC 250	PC 500	PC 1000	4	8	16	32
ImageNet class prompts										
balloon	0.20	0.53	0.51	0.37	0.49	0.61	0.38	0.51	0.49	0.57
brown bear	0.04	0.43	0.69	0.15	0.42	0.87	0.22	0.43	0.54	–
castle	0.28	0.65	0.75	0.57	0.81	0.85	0.57	0.66	0.69	0.80
cheeseburger	0.08	0.33	0.53	0.35	0.51	0.84	0.21	0.38	0.36	0.42
daisy	0.16	0.55	0.72	0.43	0.70	0.79	0.47	0.61	0.74	0.78
espresso	0.10	0.42	0.58	0.39	0.58	0.63	0.41	0.58	0.74	0.68
flamingo	0.04	0.37	0.68	0.18	0.46	0.72	0.10	0.19	0.22	0.20
golden retriever	0.10	0.50	0.74	0.25	0.42	0.59	0.42	0.50	0.53	0.84
jellyfish	0.07	0.49	0.68	0.32	0.62	0.78	0.36	0.59	0.52	0.69
lighthouse	0.14	0.48	0.80	0.34	0.58	0.68	0.32	0.33	0.40	0.47
lion	0.03	0.22	0.58	0.14	0.38	–	0.17	0.38	0.38	0.40
peacock	0.01	0.15	0.66	0.12	0.31	0.72	0.12	0.25	0.42	0.51
volcano	0.07	0.30	0.52	0.29	0.57	0.75	0.25	0.27	0.42	0.70
Open-ended prompts										
alien botanical garden	0.22	0.47	0.60	0.57	0.70	0.79	0.66	0.78	0.79	0.80
ancient library lit by fireflies	0.20	0.55	0.66	0.66	0.75	0.76	0.62	0.70	0.77	0.80
bioluminescent jellyfish	0.12	0.47	0.59	0.46	0.60	0.66	0.53	0.61	0.63	0.64
fortune teller's tent at twilight	0.19	0.50	0.66	0.39	0.59	0.69	0.46	0.54	0.65	0.60
neon samurai in rain	0.28	0.48	0.50	0.47	0.56	0.59	0.48	0.53	0.56	0.65
stained-glass cathedral interior	0.12	0.45	0.67	0.52	0.61	0.66	0.45	0.58	0.75	–
translucent koi swimming in clouds	0.05	0.42	0.68	0.39	0.62	0.78	0.50	0.60	0.72	0.77
Median	0.11	0.47	0.66	0.38	0.58	0.72	0.41	0.53	0.55	0.66

Effect of LoRA rank. Figure 18 shows each prompt rendered at $r \in \{4, 8, 16, 32\}$, with one sample per cell, picked by hand from samples ranked by Δsim. Low-rank cells stay close to the photograph, and high-rank cells move toward stylization, abstraction, or texture. In the median over prompts, the distance to vanilla and the inner-loop reward Δsim both grow with rank (Tables 5 and 6).

D Per-Prompt Natural Image Results

Table 6 gives the per-prompt values behind Figure 7, sim_0 for vanilla Stable Diffusion, and sim_0 and Δsim for our method at each rank.

Table 6: Per-prompt natural image results, restricted to our samples passing the Qwen2.5-VL relevance gate (≥ 5, same filter as Table 5). Cells in the rank columns are sim_0 / Δsim, where sim_0 is frozen-CLIP image-text cosine before adaptation (concept fidelity) and Δsim is the inner-loop reward. –: no sample passes.

	Vanilla SD	Ours (sim_0 / Δsim)			
Prompt	sim_0	r=4	r=8	r=16	r=32
ImageNet class prompts					
balloon	0.26	0.24 / 0.20	0.21 / 0.36	0.21 / 0.54	0.19 / 0.61
brown bear	0.27	0.25 / 0.20	0.20 / 0.37	0.18 / 0.55	–
castle	0.23	0.22 / 0.19	0.19 / 0.35	0.18 / 0.52	0.17 / 0.66
cheeseburger	0.26	0.23 / 0.22	0.19 / 0.39	0.21 / 0.57	0.15 / 0.70
daisy	0.27	0.23 / 0.20	0.21 / 0.36	0.19 / 0.52	0.19 / 0.59
espresso	0.24	0.22 / 0.20	0.19 / 0.37	0.21 / 0.49	0.15 / 0.67
flamingo	0.29	0.27 / 0.20	0.25 / 0.35	0.24 / 0.51	0.20 / 0.62
golden retriever	0.28	0.22 / 0.23	0.22 / 0.40	0.18 / 0.60	0.17 / 0.68
jellyfish	0.27	0.23 / 0.22	0.19 / 0.39	0.18 / 0.55	0.15 / 0.65
lighthouse	0.25	0.24 / 0.20	0.23 / 0.36	0.22 / 0.52	0.21 / 0.62
lion	0.26	0.23 / 0.19	0.21 / 0.35	0.20 / 0.53	0.21 / 0.61
peacock	0.29	0.27 / 0.23	0.24 / 0.39	0.19 / 0.56	0.16 / 0.66
volcano	0.24	0.25 / 0.21	0.24 / 0.36	0.21 / 0.54	0.22 / 0.64
Open-ended prompts					
alien botanical garden	0.27	0.20 / 0.20	0.17 / 0.36	0.15 / 0.54	0.14 / 0.62
ancient library lit by fireflies	0.28	0.20 / 0.26	0.15 / 0.44	0.12 / 0.60	0.11 / 0.70
bioluminescent jellyfish	0.28	0.17 / 0.25	0.16 / 0.43	0.10 / 0.66	0.05 / 0.82
fortune teller's tent at twilight	0.25	0.22 / 0.25	0.21 / 0.41	0.17 / 0.59	0.15 / 0.68
neon samurai in rain	0.34	0.21 / 0.26	0.17 / 0.44	0.13 / 0.60	0.13 / 0.66
stained-glass cathedral interior	0.25	0.23 / 0.24	0.19 / 0.40	0.16 / 0.55	–
translucent koi swimming in clouds	0.27	0.21 / 0.26	0.18 / 0.46	0.13 / 0.65	0.12 / 0.71
Mean	**0.27**	**0.23 / 0.22**	**0.20 / 0.39**	**0.18 / 0.56**	**0.16 / 0.66**
Median	**0.27**	**0.23 / 0.21**	**0.20 / 0.38**	**0.18 / 0.55**	**0.16 / 0.66**

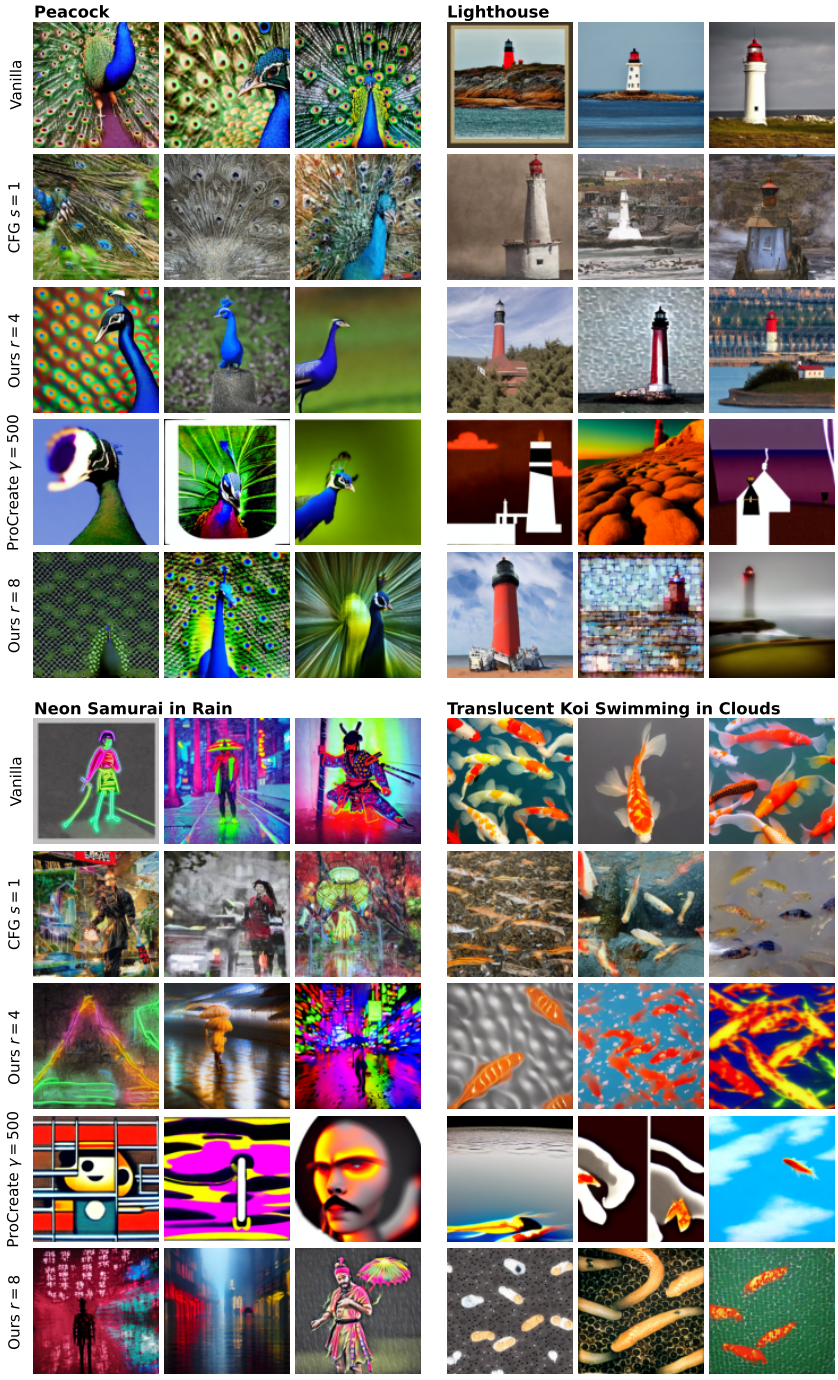


Figure 16: Samples of the diversity baselines and ours on four prompts. CFG $s{=}1$ is paired with ours at $r{=}4$ and ProCreate $\gamma{=}500$ with ours at $r{=}8$, at matched relevance rate (Table 4). Each cell is a random draw of samples with Qwen2.5-VL relevance ≥ 5. Vanilla (CFG 7.5) samples are random and ungated.

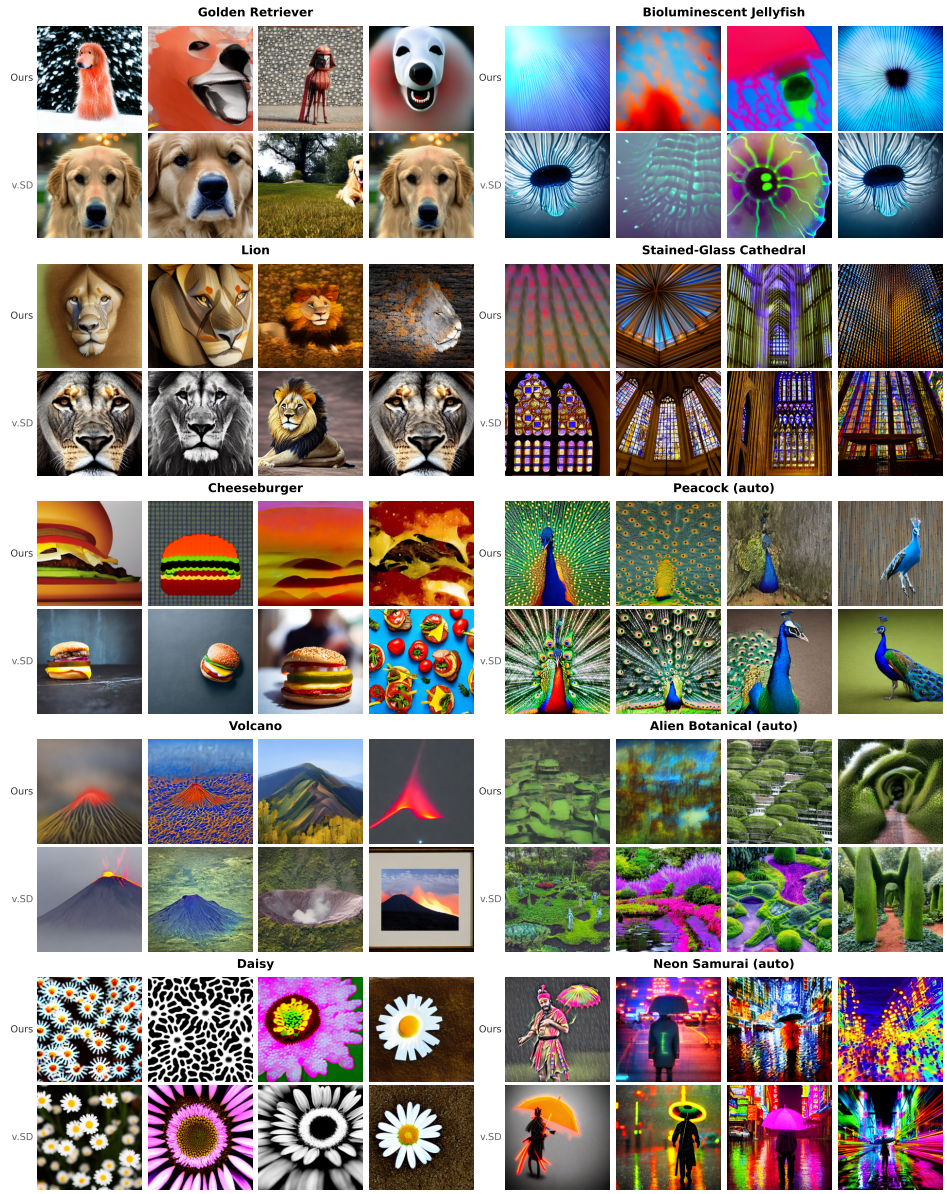


Figure 17: Each picked Creator-Appraiser sample (top of each pair) shown alongside its DINOv2-nearest neighbor (bottom of each pair) selected from 1000 vanilla Stable Diffusion samples. Top 7 prompts are the manually curated picks from Figure 6; bottom 3 are automatic picks for three of the prompts in Figure 15 (peacock, an alien botanical garden, neon samurai in rain).

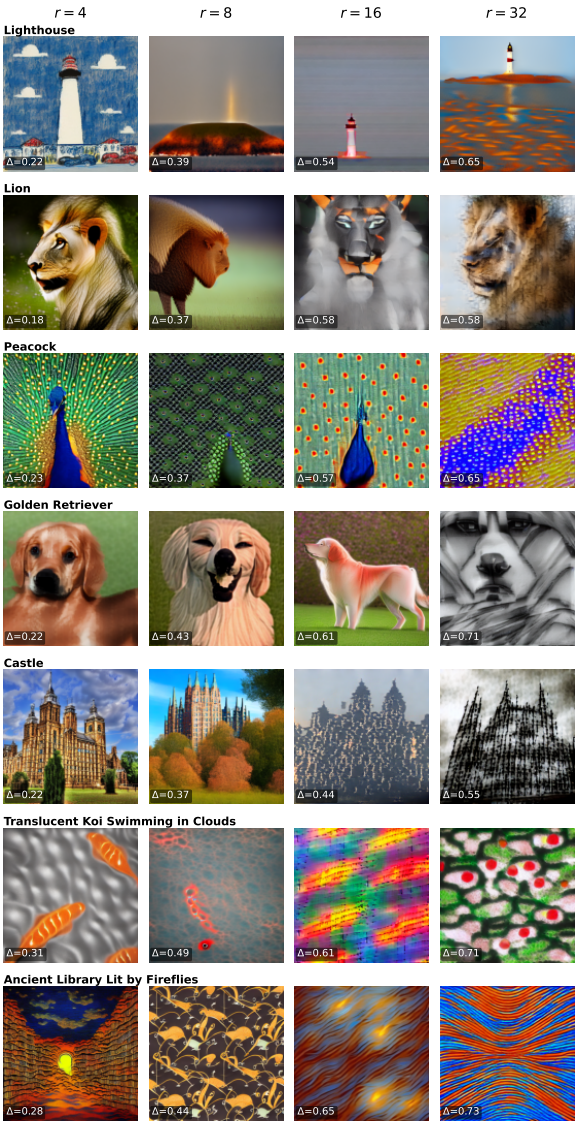


Figure 18: Effect of LoRA rank: Per prompt, one hand-picked sample at LoRA rank $r \in \{4, 8, 16, 32\}$. As rank grows the inner LoRA gains capacity to assimilate increasingly out-of-distribution outputs, so the meta-gradient is free to push the diffusion sample further from the prompt.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 3.2 state that our framework produces outputs substantially outside the sample distribution of vanilla diffusion while remaining recognizable to an off-the-shelf observer. This is supported by three independent pieces of evidence: (i) the rank-driven $\Delta\text{sim}/\text{sim}_0$/relevance trade-off in Figure 7 and Table 6; (ii) the DINOv2 reachability analysis in Section 3.2 ("Reachability beyond vanilla") and Appendix C (Table 5, Figure 17), showing our samples sit a median $4.6\times$ farther from a 1000-sample vanilla cloud than vanilla samples are from each other (VLM-relevance-filtered, $r{=}8$), with diversity baselines, selection controls, and the vanilla atypicality reference in Figure 8 and Tables 4 and 3; and (iii) qualitative picks in Figures 6, 15, and 18. The MNIST hypothesis test (Figures 3a and 4) establishes the underlying learnability signal. The introduction and Conclusion state that these claims hold for computational observers, and that agreement with human judgment is untested.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A Limitations paragraph in the Conclusion states that agreement with human judgment has not been measured, that the CLIP Appraiser does not enforce fidelity to the prompt (the relevance rate falls with adapter rank), and that the framework targets perceptual rather than functional value. The Conclusion notes that the Appraiser is memoryless across creations. Section 2.1 discusses the dependence of the induced notion of creativity on the Appraiser and its loss.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an operational framework and empirical results; it does not contain formal theorems or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithm 1 provides the full bilevel meta-gradient procedure in pseudocode. Section 2.1 details both Appraiser instantiations (autoencoder and CLIP-LoRA) including reward shaping; Section 3.2 and the top-K pipeline in Appendix B report all hyperparameters (LoRA rank, inner-loop steps and learning rate, guidance scale, denoising steps, model checkpoints). Code is released at `https://github.com/agentic-learning-ai-lab/conceptual-creativity-meta-learning`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for generation, top-K scoring, the reachability and baseline analyses, and figure generation is released at `https://github.com/agentic-learning-ai-lab/conceptual-creativity-meta-learning`. All datasets used (MNIST, Omniglot for reference stimuli, ImageNet validation labels) and all generative and evaluation models (Stable Diffusion v1.5, CLIP ViT-L/14, Qwen2.5-VL-7B) are publicly available pretrained checkpoints.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters and architectural details are specified in Section 2.1.1 (Autoencoder Appraiser) and Section 2.1.2 (CLIP Appraiser), with experiment-specific settings in the Setup paragraph of Section 3.2 and the top-K pipeline in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 7 reports per-rank means with ± 1 standard-deviation bands across prompts; Tables 6 and 5 give per-prompt entries (means over $N = 100$ generations per prompt for our method, $N = 1000$ for the vanilla samples). The MNIST scatter plots (Figs. 3a, 4, 5) plot individual generations rather than aggregate error bars, so the within-category dispersion is directly visible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Natural-image generation runs on a single H200 or L40S GPU per sweep cell (12–15 minutes per prompt-rank combination at 100 samples, 50 inference steps, 10 inner-loop steps). The sweep across 20 prompts $\times$ four LoRA ranks takes about 17 GPU-hours. Measured on one H200, our method takes 7.7 s per image and vanilla Stable Diffusion 0.86 s (Appendix C). The baselines in Appendix Table 4 run on the same hardware. Top-K scoring with Qwen2.5-VL-7B in fp16 takes $\sim$5 minutes per generation directory on an H200 ($\sim$10 GPU-hours total across all dirs). MNIST experiments run on a single workstation GPU (1080 Ti) in under an hour each.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and pretrained models, does not involve human subjects, and does not raise privacy or fairness concerns specific to the proposed method.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our framework allows generative models to produce stylistically novel outputs from existing prompts; potential positive impacts include creative-tool augmentation for artists and designers, and a more honest computational handle on what makes an output "creative" as opposed to merely deviant. Potential negative impacts mirror those of underlying generative models (e.g., production of misleading imagery or stylistically-confounded copyright-adjacent material); the framework does not introduce new misuse vectors beyond those of the Stable Diffusion v1.5 backbone it builds on.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release new pretrained generative models or scraped datasets. The framework uses publicly available Stable Diffusion v1.5 and CLIP weights without fine-tuning either; the only learned object per generation is a small (rank ≤ 32) per-image LoRA adapter on a text projection that is discarded after sampling.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets are credited via citations: Stable Diffusion v1.5 [36] (CreativeML Open RAIL-M), CLIP ViT-L/14 [33] (MIT), Qwen2.5-VL-7B [2] (Apache 2.0), MNIST and Omniglot (research use), ImageNet validation labels for class-conditioned prompts (non-commercial research). All are used within their permitted research contexts.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new datasets or pretrained models. The contribution is the framework, the algorithm (Algorithm 1), and the curation/evaluation code; the latter is released with documentation at the code URL given above.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen2.5-VL-7B-Instruct is used as a relevance scorer in the automatic curation pipeline reported in Appendix B, judging concept-presence on a 0–10 scale. The scoring instructions, gating threshold, and post-processing are summarized there, and the exact

prompt is in the released code. The VLM serves only as an evaluation measure and relevance gate; it is not part of the framework, the algorithm, or the MNIST experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.