Shaping Useful Noise: Energy Distributions Predict Visual Pretraining Quality

Ching Lam Choi[1] **Antonio Torralba** **Phillip Isola** **Stefanie Jegelka**[1,2]
[1]CSAIL, Department of EECS, Massachusetts Institute of Technology
[2]School of CIT, MCML, MDSI, Technical University of Munich
{chinglam, torralba, phillipi}@mit.edu, stefanie.jegelka@tum.de

Abstract

Synthetic data can scale visual pretraining, but current selection criteria only partially explain which datasets train strong representations. We propose using the *energy profile* of a dataset—the distribution of log-probability levels it occupies under a normalised real-image density model—as an audit and curation principle for synthetic data. To operationalise this view, we introduce *Energy-Level Distribution Distance (*ELD*)*, which compares real and synthetic energy distributions. Real images span sparse high-likelihood scenes and dense low-likelihood textures, so useful synthetic data should cover this energy range rather than collapse to likely but low-variation samples. Across diverse synthetic datasets, ELD exposes mismatches missed by FID and predicts pretraining performance beyond FID, LPIPS variation, and feature-space recall, including across CNN and ViT self-supervised settings. We then use ELD to resample and mix synthetic sources, matching the real-image energy histogram while preserving diversity and improving downstream transfer.

1 Introduction

Synthetic visual pretraining turns data choice into a distributional design problem: models can be trained on simulations, procedural images, generator samples, or mixtures thereof, each inducing different visual statistics. Yet current selection criteria often reduce this choice to sample realism or feature-space similarity. These criteria are useful, but they compress datasets into distances, neighbourhoods or low-order feature summaries, and do not ask where samples lie in the probability landscape of real images. This missing axis matters because natural images do not occupy a single typical likelihood level. Normalised density models reveal a broad, content-dependent energy distribution, where sparse simple scenes and dense textured scenes occupy different likelihood regimes (Guth et al., 2026; Kadkhodaie et al., 2024). This echoes classical natural-image statistics, which show that real images have structured spectral, wavelet and sparsity properties far from white noise (Burton & Moorhead, 1987; Field, 1987; Ruderman, 1997; Simoncelli, 2005; Portilla & Simoncelli, 2000). A synthetic dataset may therefore match real feature moments while exposing a learner to the wrong likelihood regimes; selecting only the most likely samples may likewise remove complex variation that real data contains. We ask whether synthetic data should be selected and designed to match the *distribution of likelihood levels* occupied by real images.

1.1 Our Contributions

We make four contributions around a common claim: beyond visual realism or feature-space similarity, *the energy shape of synthetic data is a useful predictor and design principle for visual pretraining.*

First, we formalise this design axis through Energy-Level Distribution Distance (ELD), together with a percentile variant, which compares real and synthetic datasets through their normalised real-image

Preprint.

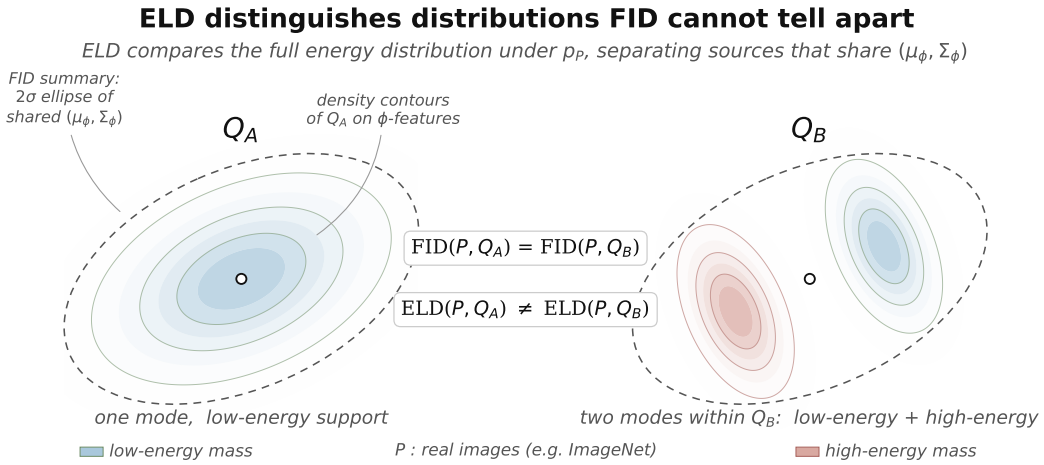

Figure 1: **ELD distinguishes synthetic pretraining datasets FID cannot.** Self-supervised pretraining benefits when synthetic data spans the full likelihood spectrum of real images; sources with identical FID can carry very different energy distributions, which ELD detects.

log-probability distributions. Unlike mean likelihood, ELD measures the full energy profile of a dataset. It penalises synthetic data that drifts into low-likelihood, noise-like regions, as well as data that collapses towards high-likelihood but low-variation samples. This makes ELD complementary to feature-space metrics such as FID, LPIPS variation and recall, which measure moments, perceptual distances or coverage rather than probability-level structure.

Second, we use ELD to audit a broad suite of procedural, statistical, feature-visualisation and generator-based synthetic datasets. The audit reveals three recurring regimes: noise-like datasets that sit in low-likelihood regions, over-concentrated datasets that collapse around high-likelihood but low-variation samples, and broad-but-shifted datasets that span many energy levels while missing the real profile. These regimes expose why feature-space metrics can be incomplete: two datasets may have similar FID, LPIPS variation or recall while presenting very different likelihood structure to the learner. Third, we show that this structure is predictive, not merely descriptive: ELD explains pretraining utility beyond standard metrics, including across CNN and ViT self-supervised settings. Finally, we turn the audit into an intervention. By resampling within synthetic pools and mixing across generators, we match the real-image energy histogram without sacrificing feature diversity, yielding synthetic training sets with stronger downstream transfer.

1.2 Related Work

Synthetic data for visual pretraining. Synthetic visual data has been explored through several generation mechanisms: simulation and domain randomisation for transfer to real environments (Tobin et al., 2017), fractal and procedural image datasets for pretraining without photographs (Kataoka et al., 2020), structured noise and statistical image models for contrastive representation learning (Baradad Jurjo et al., 2021), and text-to-image generators for large-scale synthetic representation learning (Tian et al., 2023). These works establish a rich design space for synthetic pretraining data. Our focus is not to introduce a new generator, but to identify a distribution-level property that predicts which generated datasets are useful across modern self-supervised objectives.

Distribution metrics for generated data. Most metrics compare real and generated distributions in a learned feature space. FID fits Gaussian approximations to Inception features (Heusel et al., 2017); KID uses a kernel distance (Bińkowski et al., 2018); precision and recall estimate fidelity and coverage (Kynkäänniemi et al., 2019; Naeem et al., 2020); and LPIPS measures perceptual similarity with deep features (Zhang et al., 2018). These metrics capture fidelity and diversity, but recent work shows their sensitivity to feature choice and Gaussian assumptions (Jayasumana et al., 2024). ELD is orthogonal to these criteria: it compares probability-level structure under a real-image density model rather than distances or moments in representation space.

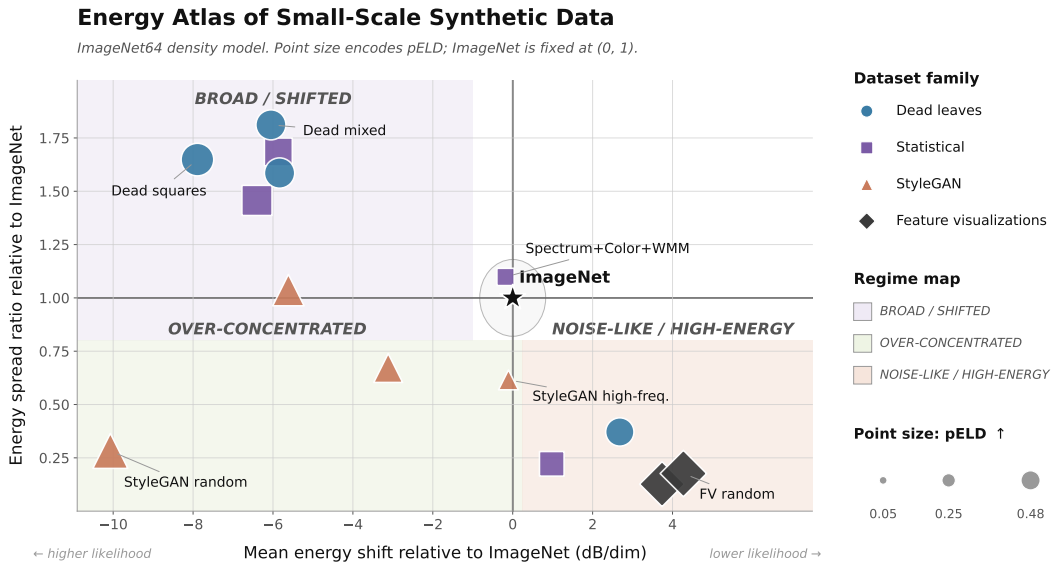

Figure 2: **Small-scale energy audit.** Synthetic datasets scored under a grayscale ImageNet64 density model. Axes show mean energy shift and spread relative to ImageNet, while point size denotes pELD. The atlas provides a coarse diagnostic view of the regimes expanded in Figure 3: noise-like sources are shifted toward high-energy regions, over-concentrated sources occupy too narrow a likelihood band, and broad/shifted sources cover a wider range but remain displaced from the real energy profile.

Image likelihoods and energy structure. Energy-based models describe data by negative log density, but calibrated image likelihoods are hard to estimate (Hinton et al., 1986; LeCun et al., 2006; Song & Kingma, 2021). Normalising flows give exact likelihoods through invertible transformations (Rezende & Mohamed, 2015; Dinh et al., 2014, 2017), while diffusion and score-based models estimate image distributions through denoising dynamics (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Recent normalised density estimators enable direct analysis of natural-image likelihoods; Guth et al. (2026) show that real-image log probabilities are broad, content-dependent, and approximately Gumbel-like. We use this likelihood-level view as a diagnostic and curation principle: it identifies which synthetic datasets are mismatched and how to reshape them.

2 Energy Structure of Natural and Synthetic Images

A useful synthetic pretraining distribution should reproduce more than the visual appearance of real images. We focus on one structural axis: the distribution of likelihood levels assigned to a dataset by a normalised real-image density model. This energy distribution is rich in natural images; it is not determined by FID or other moment-based feature-space metrics, and separates synthetic sources into three regimes with distinct measurable signatures.

2.1 The Energy Distribution of Natural Images

Let P be a real-image distribution and p_P a normalised density model fitted to P. The *real-image energy* $E_P(x) = -\log p_P(x)$ scores how typical x is under the model: low for images p_P assigns high likelihood, high for images p_P assigns low likelihood. The pushforward $(E_P)_{\#}P$ records how often real images occur at each likelihood level. This energy distribution is broad: Guth et al. (2026) show that on ImageNet it is approximately Gumbel-like, with sparse simple scenes at the low-energy/high-likelihood end and dense textured scenes forming a high-energy/low-likelihood tail, consistent with the statistical inhomogeneity of natural images (Field, 1987; Ruderman, 1997; Simoncelli, 2005; Portilla & Simoncelli, 2000).

The implication for synthetic pretraining is concrete. A useful synthetic distribution should expose the learner to the same range of likelihood levels as real data, not merely maximise average likelihood. Pushing $\mathbb{E}_Q[\log p_P]$ upward can collapse Q to the mode of p_P; ignoring likelihood can place mass

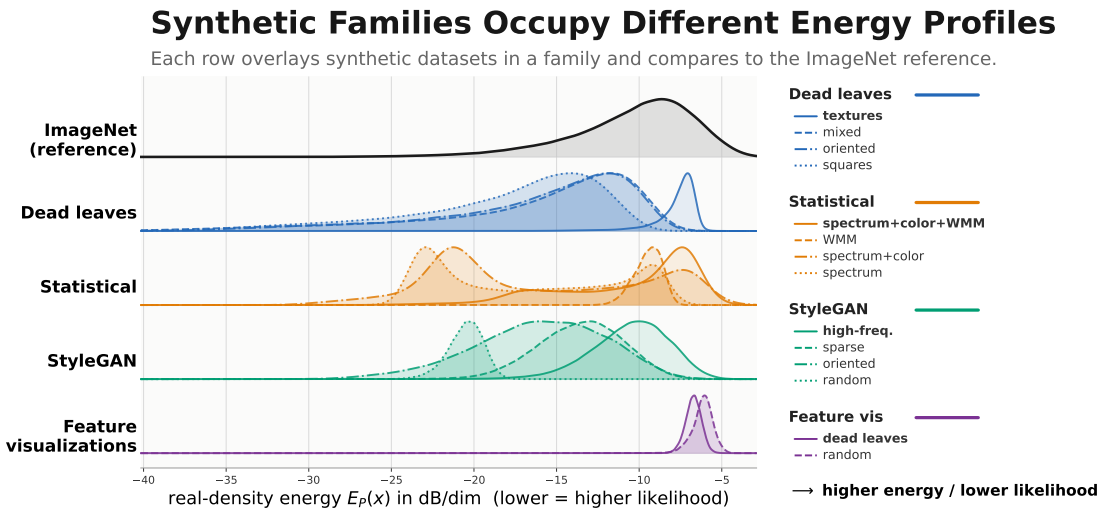

Figure 3: **Energy mismatch is structured.** Synthetic sources differ in how they depart from ImageNet: some collapse into narrow high-likelihood bands, others extend into low-likelihood tails, and others remain broad but displaced. Effective synthetic pretraining may require matching the full probability-level profile of natural images, not simply maximising likelihood or feature realism.

arbitrarily far in the high-energy tail. Either failure appears as a mismatch in $(E_P)_{\#}Q$. The relevant comparison is therefore not a scalar summary such as mean likelihood, but the full energy profiles $(E_P)_{\#}P$ and $(E_P)_{\#}Q$, which Section 3.2 measures.

2.2 Feature-Moment Metrics Are Energy-Blind

A natural question is whether existing distributional metrics already capture energy matching. For feature-moment metrics such as FID, the answer is no for structural reasons. FID compares Gaussian fits to $\phi_{\#}P$ and $\phi_{\#}Q$ for a fixed feature map ϕ (Heusel et al., 2017), and therefore depends on Q only through the first two moments of $\phi(x)$. It can distinguish feature-moment mismatch, but not how probability mass is arranged across likelihood levels under p_P: two datasets can look identical to FID while one collapses near the mode and another lies in the tail.

Proposition 2.1 (FID-level sets contain distinct energy profiles)**.** *Fix ϕ, E_P, and a reference distribution R with feature moments (μ, Σ). Let $\Pi_\phi(\mu, \Sigma)$ denote the set of image distributions whose ϕ-features have mean μ and covariance Σ, and disintegrate any $S \in \Pi_\phi(\mu, \Sigma)$ as $S(dx) = S_z(dx)\,\nu(dz)$ with $z = \phi(x)$.*

Suppose S admits measurable fibre selections $x^-(z), x^+(z) \in \operatorname{supp} S_z$ whose induced energy laws differ under ν. Then there exist $Q^-, Q^+ \in \Pi_\phi(\mu, \Sigma)$ with identical FID to R but different energy pushforwards. Moreover, if on a feature set Z of mass α these selections satisfy $E_P(x^-(z)) \le a$, $E_P(x^+(z)) \ge a + \Delta$, and agree off Z, then

$$W_1\big((E_P)_{\#}Q^-, (E_P)_{\#}Q^+\big) \ge \alpha\Delta. \tag{1}$$

The proof (Appendix A) moves mass within fibres of ϕ, preserving feature values, moments, and FID while shifting probability mass across energy levels. Feature-space diversity and coverage metrics can inherit similar blind spots, so energy matching provides an independent evaluation axis.

2.3 A Diagnostic Taxonomy of Synthetic Energy Mismatch

If FID does not determine energy matching, the remaining question is empirical: do synthetic datasets differ along this axis? Figure 2 visualises representative sources scored under a shared real-image density model p_P. Let μ_P, σ_P denote the mean and standard deviation of E_P on real data, and μ_Q, σ_Q the corresponding quantities for a synthetic source. Three regimes recur. *Noise-like* datasets have $\mu_Q - \mu_P$ large and positive relative to σ_P, placing mass in the high-energy tail: the real model rates these images as implausible. *Over-concentrated* datasets have small σ_Q/σ_P, often with $\mu_Q \le \mu_P$:

samples may be individually plausible, but occupy too narrow a slice of natural-image structure. *Broad/shifted* datasets have substantial mismatch in either spread or location, with $|\mu_Q - \mu_P|$ and/or σ_Q/σ_P far from the real reference. Mixed cases interpolate between these regimes. Figure 3 shows the same regimes at the level of full energy distributions, making the shift, compression, and tail mismatches visible beyond the two summary statistics.

These diagnostics depend on the calibration of p_P; Section 3.2 introduces a percentile-rank variant that removes this dependence. By Proposition 2.1, datasets in different regimes can have similar FID despite different energy arrangements, and Section 4.2 confirms this empirically. These regimes show why "realism" is incomplete for synthetic pretraining: a dataset can be too unlikely, too concentrated near high-likelihood regions, or mismatched across likelihood levels. The target is the full real-image energy distribution, not a scalar likelihood or feature-moment score.

3 ELD: Energy-Level Distribution Distance

Section 2.2 shows that feature-moment metrics do not determine a dataset's energy profile. We now define a metric that compares datasets along this axis directly.

3.1 Definition and Estimation

Given a normalised real-image density model p_P with energy $E_P(x) = -\log p_P(x)$, the *Energy-Level Distribution Distance* between real and synthetic distributions P, Q is

$$\mathrm{ELD}(P, Q) \;=\; W_1\big((E_P)_{\#}P,\, (E_P)_{\#}Q\big), \tag{2}$$

the Wasserstein-1 distance between their one-dimensional energy distributions. In one dimension, W_1 equals the integral distance between empirical CDFs (Villani, 2021), making ELD simple to estimate from sorted energy scores. We use W_1 rather than KL because it remains well-defined under support mismatch and returns an interpretable nat-scale value.

Raw energies depend on the calibration of p_P, so we also use a percentile version. Let F_P denote the CDF of real-image energies $E_P(x)$ for $x \sim P$. We replace each energy by its real-energy rank $F_P(E_P(x)) \in [0, 1]$. By construction, $(F_P \circ E_P)_{\#}P = \mathrm{Unif}(0, 1)$, so

$$\mathrm{pELD}(P, Q) \;=\; W_1\big(\mathrm{Unif}(0, 1),\, (F_P \circ E_P)_{\#}Q\big). \tag{3}$$

pELD is bounded, invariant to monotone recalibrations of p_P, and vanishes exactly when synthetic energies match the real energy distribution. Unless otherwise stated, we report pELD as ELD; both variants compare the same energy ordering, but pELD removes dependence on absolute calibration.

We instantiate p_P with a diffusion-based density estimator, which scales to ImageNet resolution and provides tractable image log-likelihood estimates (Ho et al., 2020; Kingma et al., 2021). Once energies are computed, ELD requires only empirical CDFs; Section 4.2 shows that a few thousand samples per distribution suffice for stable rankings and tests robustness to the density model. The main cost is energy evaluation, amortised across all synthetic datasets scored under the same p_P.

3.2 Interpreting ELD

ELD turns the qualitative regimes of Section 2.3 into a single distributional discrepancy. If a synthetic dataset is shifted into the high-energy tail, compressed around high-likelihood modes, or broad but displaced from the real profile, its empirical energy CDF departs from that of real images. The Wasserstein distance integrates these departures across the full likelihood axis, so the metric penalises datasets that are too unlikely, too concentrated, or likely at the wrong frequencies. This is the distinction missed by mean likelihood: ELD does not reward making every sample as likely as possible, but rather matching how real images populate likelihood levels.

This interpretation also clarifies the role of ELD. Rather than replacing semantic or feature-space coverage metrics, ELD isolates a complementary question: whether a dataset populates the same likelihood levels as real images under a shared density model. Two datasets may match in energy profile while differing in semantics or feature diversity; conversely, they may look similar to feature-moment metrics while occupying very different regions of the real-image likelihood axis. We therefore report ELD alongside FID, LPIPS variation, and recall in Section 4.2. Figure 3 illustrates the measured object: when scored by the same ImageNet density model, synthetic families induce distinct energy profiles with different shifts, spreads, and tail structure relative to ImageNet.

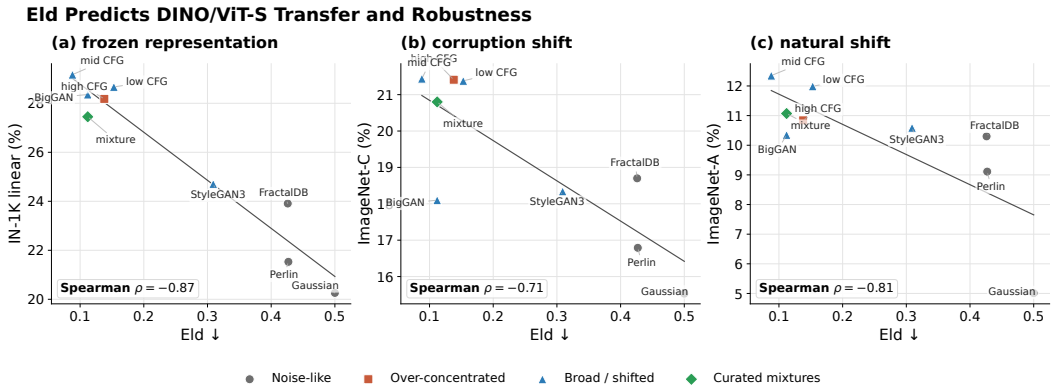

Figure 4: **ELD predicts DINO/ViT-S downstream quality.** Across the self-supervised synthetic pretraining suite, lower ELD is associated with stronger frozen ImageNet-1K linear-probe accuracy and better robustness on ImageNet-C and ImageNet-A. Points are colored by the energy-regime taxonomy, showing that high-ELD noise-like sources occupy the weak-transfer end of the trend, while lower-ELD diffusion and mixture sources cluster among the strongest representations.

4 Energy Shape Predicts Pretraining Utility

Sections 2.2 and 3.2 argue that energy structure is an axis FID does not see and that ELD measures it. We now ask whether this axis matters: across a broad suite and self-supervised recipes, ELD predicts downstream accuracy beyond FID, LPIPS variation, and recall.

4.1 Experimental Setup

We ask two questions: does ELD predict how well a synthetic dataset trains visual representations, and does it explain variance beyond FID, LPIPS variation, and recall? We evaluate this in two complementary pretraining settings. The first is a self-supervised setting using DINO-style ViT pretraining (Caron et al., 2021). The second follows the setting of Baradad Jurjo et al. (2021), where synthetic datasets are used directly as pretraining sources under the original CNN/contrastive evaluation protocol. In both settings, we score each synthetic dataset under the same metric suite and regress downstream accuracy on these metrics. Appendix B.2 visualizes samples from the corresponding DINO/ViT-S and Learning-to-See synthetic sources.

Synthetic pretraining data. The suite spans the regimes of Section 2.3: procedural and noise-like sources from Baradad Jurjo et al. (2021), including dead leaves, statistical image models, feature visualizations, and untrained StyleGAN variants; classical synthetic pretraining sources such as FractalDB, Perlin noise, and Gaussian noise; generator-based sources including StyleGAN3, BigGAN, and diffusion/latent-diffusion samples at varied guidance or sampling settings; and curated mixtures such as StableRep and uniform mixtures (Baradad Jurjo et al., 2021; Kataoka et al., 2022; Karras et al., 2021; Brock et al., 2019; Rombach et al., 2022; Tian et al., 2023).

Pretraining recipes. Our first recipe is DINO with ViT-S under fixed hyperparameters across datasets (Caron et al., 2021). To connect with earlier synthetic-data pretraining work, we also run the "Learning to See by Looking at Noise" protocol using its corresponding CNN-based recipe and synthetic dataset families (Baradad Jurjo et al., 2021).

Evaluation. For the self-supervised setting, we evaluate ImageNet-1K representation quality using SimCLR-style linear probing and fine-tuning (Chen et al., 2020). For the Learning-to-See setting, we follow the original ImageNet100 evaluation protocol of Baradad Jurjo et al. (2021). In both settings, we additionally evaluate distribution shift on ImageNet-C and ImageNet-A (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021); for Learning-to-See, these shift evaluations are restricted to the ImageNet100 class subset and resized to the encoder input resolution. For each pretraining dataset we compute ELD against ImageNet alongside FID, LPIPS, and Improved Recall (Heusel et al., 2017; Zhang et al., 2018; Kynkäänniemi et al., 2019). Table 1 reports DINO/ViT-S, while Table 2

Dataset	Data metrics				DINO/ViT-S downstream			
	ELD↓	FID↓	LPIPS↑	Recall↑	IN-1K Lin.%↑	IN-1K FT%↑	IN-C%↑	IN-A%↑
Noise-like								
FractalDB	**.426**	1254	.260	.018	**23.91**	**56.66**	**18.70**	**10.30**
Perlin noise	.427	1922	.208	.000	21.53	55.38	16.79	9.11
Gaussian noise	.500	2145	.237	.000	20.26	52.26	15.53	5.01
Over-concentrated								
diffusion (high CFG)	**.138**	481	.342	.020	**28.18**	**59.34**	**21.41**	**10.84**
Broad / shifted								
StyleGAN3	.309	1148	.305	.001	24.71	57.49	18.34	10.59
BigGAN	.112	566	.303	.101	28.36	58.97	18.10	10.35
diffusion (mid CFG)	**.088**	459	.338	.030	**29.17**	59.57	**21.44**	**12.35**
diffusion (low CFG)	.153	1129	.297	.022	28.67	**59.60**	21.38	11.99
Curated mixtures								
uniform mixture	**.112**	616	.294	.023	**27.45**	**57.61**	**20.80**	**11.07**

Table 1: **Self-supervised setting.** Dataset metrics and DINO/ViT-S accuracy; metrics are computed against ImageNet, and evaluation reports ImageNet-1K linear/FT plus ImageNet-C/A robustness.

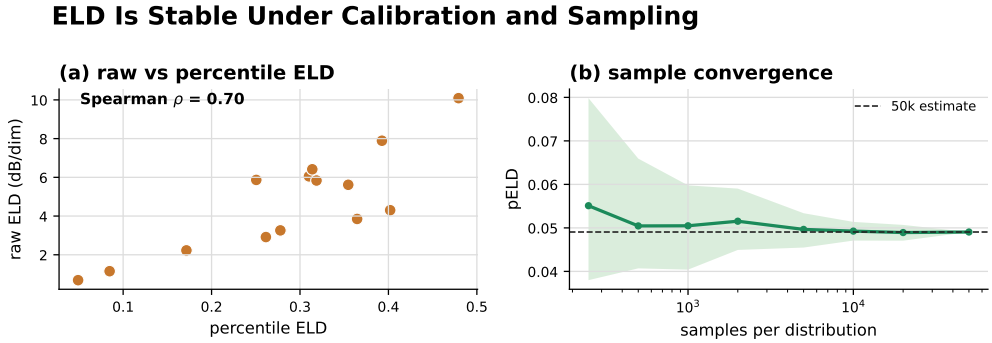

Figure 5: **ELD is robust to estimation choices.** Raw and percentile distances agree with a Spearman coefficient of 0.7; subsampling shows stable estimates after a few thousand scored samples.

reports the Learning-to-See setting. We use DINO/ViT-S as the modern self-supervised protocol and Learning-to-See as a complementary small-scale replication with classical synthetic pretraining.

4.2 ELD Predicts Downstream Quality Beyond FID

We next test whether probability-level mismatch predicts pretraining utility in a modern self-supervised setting. Table 1 reports DINO/ViT-S models pretrained on each synthetic source, with ImageNet-1K linear probing, ImageNet-1K fine-tuning, and ImageNet-C/A robustness. Across this suite, ELD is consistently negatively associated with downstream performance: lower ELD corresponds to higher frozen ImageNet-1K linear accuracy (Spearman $\rho = -0.87$, $p = 0.002$), higher ImageNet-1K fine-tuning accuracy ($\rho = -0.78$, $p = 0.014$), and better robustness on ImageNet-C ($\rho = -0.71$, $p = 0.032$) and ImageNet-A ($\rho = -0.81$, $p = 0.008$). These results support the directional prediction of Section 2.3: sources whose energy profiles depart from the real-image likelihood distribution yield weaker transferable representations. Figure 4 visualises the trend, with regimes organizing the scatter from high-ELD noise-like sources at the weak-transfer end to lower-ELD diffusion and mixture sources among the strongest models.

We also evaluate ELD in the classical synthetic-pretraining setting of Baradad Jurjo et al. (2021). Table 2 reports the small-scale synthetic families under the original ImageNet100 transfer protocol, augmented with ImageNet-C/A evaluations. This benchmark is heterogeneous by construction: dead leaves, statistical models, untrained StyleGANs, and feature visualizations differ not only in

	Data metrics				Learning-to-See downstream			
Dataset	ELD ↓	FID ↓	LPIPS ↑	Recall ↑	IN Lin.% ↑	IN FT% ↑	IN-C% ↑	IN-A% ↑
Dead leaves								
dead-leaves squares	.393	1480	.316	.005	38.60	67.52	27.23	3.24
dead-leaves oriented	.319	1390	.299	.035	40.90	68.48	25.12	**3.97**
dead-leaves mixed	.310	1310	.301	.074	40.84	**68.86**	25.00	3.82
dead-leaves textures	**.261**	1340	.241	.084	**41.00**	68.20	**27.93**	3.83
Statistical image models								
spectrum	.314	1560	.159	.003	30.70	62.20	24.57	3.68
WMM	.172	1520	.152	.003	34.08	63.30	16.90	3.53
spectrum+color	.251	1290	.258	.001	37.88	65.18	27.82	4.12
spectrum+color+WMM	**.049**	1210	.228	.010	**38.84**	**65.62**	**30.28**	**4.56**
Untrained StyleGAN								
StyleGAN random	.479	1420	.279	.000	35.30	64.82	29.10	**4.41**
StyleGAN high-freq.	**.085**	1320	.268	.002	**39.62**	**66.21**	**29.84**	4.38
StyleGAN sparse	.278	1240	.305	.001	38.76	66.16	27.23	3.24
StyleGAN oriented	.355	903	.272	.007	39.08	65.84	27.70	3.38
Feature visualizations								
feature-vis random	.402	1590	.252	.000	35.26	66.22	22.89	2.65
feature-vis dead leaves	**.365**	1400	.223	.001	**38.70**	**68.00**	**32.04**	**4.71**

Table 2: **Learning-to-See synthetic-pretraining setting.** Dataset metrics and downstream accuracy for the small-scale families from Baradad Jurjo et al. (2021). Evaluation follows the original ImageNet protocol and adds ImageNet-C/A robustness with the same representations.

energy profile but also in construction mechanism and semantics. Comparing across heterogeneous synthetic sources, the pooled correlations remain directionally consistent: lower ELD is associated with higher ImageNet-A accuracy (Spearman $\rho = -0.26$) and weakly with higher ImageNet-C accuracy ($\rho = -0.15$). We note that ImageNet-A is evaluated through the ImageNet100 classifier, so classes outside this 100-class subset incur errors and render the absolute accuracies conservative.

The within-family comparisons are particularly informative. Among dead-leaves variants, lower ELD tracks stronger early ImageNet transfer ($\rho = -0.80$ for linear accuracy), with the lowest-ELD texture variant achieving the best linear accuracy and ImageNet-C accuracy in the group. Among statistical image models, ELD ranks both ImageNet linear and fine-tuned transfer strongly ($\rho = -0.80$ for each), and the best energy match, spectrum+color+WMM, is also the strongest source on ImageNet linear transfer, fine-tuning, ImageNet-C, and ImageNet-A. The untrained StyleGAN family shows the same pattern for ImageNet transfer: lower ELD is associated with stronger linear accuracy ($\rho = -0.80$) and fine-tuned accuracy ($\rho = -1.00$), with the high-frequency variant attaining the best transfer scores. For the two feature-visualization sources, the lower-ELD dead-leaves variant is better on every downstream metric. Thus, although ELD is not intended to collapse this mixed suite into a single universal ordering, it provides a consistent directional signal when comparing synthetic sources within a shared family.

Together, the two experimental settings support the regime picture of Section 2.3. ELD does not subsume all factors governing synthetic pretraining: semantic coverage, augmentation compatibility, and feature diversity remain important. Instead, it isolates a complementary axis of mismatch, namely whether a synthetic source places probability mass at the likelihood levels occupied by real images. The DINO/ViT-S results show that this axis predicts transfer and robustness in a modern self-supervised protocol, while the Learning-to-See results show that the same signal remains interpretable in a classical small-scale synthetic-pretraining setting, especially within matched generator families. This is the role of ELD in our analysis: not a universal replacement for feature-space metrics, but a probability-level diagnostic that explains a structured component of downstream variation.

Estimation robustness. Figure 5 verifies that the reported percentile version is not an artefact of calibration or sample size. Raw energy-space distance and percentile ELD are strongly aligned across datasets, indicating that rank calibration preserves the relevant ordering while removing dependence

on the absolute scale of the density model. Subsampling shows that ELD estimates stabilise after a few thousand scored examples, well below the sample sizes used in the main experiments.

5 Energy-Aware Curation for Synthetic Data

ELD suggests a constructive use beyond evaluation: synthetic data can be *shaped* to match the probability-level profile of real images. Natural images do not concentrate at a single likelihood level; they occupy a broad energy spectrum containing both simple high-likelihood scenes and dense low-likelihood structure (Guth et al., 2026; Field, 1987; Ruderman, 1997; Simoncelli, 2005). This suggests that the naive approach to density scoring—filtering for the highest-likelihood samples—may be a misguided intervention. The goal is not to keep only the most probable synthetic samples, but to make synthetic pretraining data populate real-image likelihood levels in the right proportions.

Because real samples are uniform in real-energy percentile space, a simple curation rule is to resample synthetic data so that its percentile histogram is approximately uniform. In practice, we score a candidate pool under p_P, bin samples by their real-energy percentiles, and draw a balanced training stream across bins while preserving within-bin diversity. For mixtures, the same idea becomes a small convex design problem: choose source weights whose combined energy histogram best matches the real one. Thus the audit becomes prescriptive: over-concentrated sources should be supplemented with higher-energy structure, noise-like sources should be pulled toward plausible mid-energy regions, and broad shifted sources should be used only where they fill missing likelihood bands.

ELD further gives a synthetic data mixture curation rule rather than only a score. Given a synthetic pool with cached energies and features, we may select samples to match the real energy histogram while preserving diversity within each energy band. This changes only which generated images are used: no generator is retrained, no labels are introduced, and no downstream information is used. This seeks to reintroduce under-covered likelihood bands, not merely keeping the most likely samples. More broadly, ELD reframes synthetic-data construction as distributional engineering. A curated dataset is a measure whose pushforward under E_P should approximate the real pushforward $(E_P)_{\#}P$, while still satisfying feature-coverage or diversity constraints. This turns common design choices—which samples to keep, which generators to mix, which prompts or guidance scales to use—into ways of moving mass along the real-image energy axis. The target is not maximal realism or likelihood, but the full probability-level profile of natural images.

6 Conclusion

Discussion and Limitations. ELD makes synthetic data inspectable at the level of likelihood structure: rather than asking only whether samples look realistic or match feature moments, it asks which parts of the real-image probability landscape they occupy. This view treats the density model as part of the experimental specification; different choices of p_P may emphasise different domains or image statistics, so ELD should be reported together with its scorer. The metric complements FID, recall, and LPIPS variation by measuring probability-level occupancy, while feature-space metrics capture realism, diversity, and coverage in representation space. Our experiments focus on visual pretraining, where energy structure predicts representation quality; objectives such as supervised-label efficiency, aesthetic fidelity, or human preference may require additional criteria. Finally, our analysis is image-only. CLIP-style multimodal pretraining introduces text and image-text pairing distributions, which may require matching both visual and linguistic structure.

Synthetic data is not characterised by visual realism or feature-space similarity alone; its position in the real-image probability landscape also matters. ELD provides a way to measure this axis by comparing the distribution of likelihood levels induced by real and synthetic data under a shared density model. We show that synthetic sources occupy distinct energy regimes that are not captured by feature-moment metrics, and that these regimes are predictive of downstream pretraining quality. Since the mismatch is structured, ELD is also prescriptive: it identifies likelihood bands that are missing, overrepresented, or shifted, enabling resampling and mixing strategies that better match the energy profile of natural images. More broadly, this points to a design view of synthetic data: generators, filters, and mixtures can be selected not only for realism or diversity, but for how they populate the real-image likelihood spectrum.

References

Baradad Jurjo, M., Wulff, J., Wang, T., Isola, P., and Torralba, A. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34:2556–2569, 2021. 2, 6, 7, 8

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 2

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm. 6

Burton, G. J. and Moorhead, I. R. Color and spatial structure in natural scenes. *Applied optics*, 26(1): 157–170, 1987. 1

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. 6

Castaing, C. and Valadier, M. Convex analysis and measurable multifunctions. *Lecture Notes in Mathematics*, 580, 1977. 13

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html. 6

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 3

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx. 3

Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987. 1, 3, 9

Guth, F., Kadkhodaie, Z., and Simoncelli, E. P. Learning normalized image densities via dual score matching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=wtYcS4kxpF. 1, 3, 9

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm. 6

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021. 6

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf. 2, 4, 6

Hinton, G. E., Sejnowski, T. J., et al. Learning and relearning in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317):2, 1986. 3

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf. 3, 5

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, June 2024. 2

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0. 1

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 6

Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., and Satoh, Y. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Zhang, X., Martinez-Noriega, E. J., Inoue, N., and Yokota, R. Replacing labeled real-image datasets with auto-generated contours. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21232–21241, 2022. 6

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 5

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 2, 6

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020. 2

Portilla, J. and Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000. 1, 3

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015. 3

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 6

Ruderman, D. L. Origins of scaling in natural images. *Vision research*, 37(23):3385–3398, 1997. 1, 3, 9

Simoncelli, E. P. 4.7 statistical modeling of photographic images. *Handbook of Video and Image Processing*, 9, 2005. 1, 3, 9

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015. 3

Song, Y. and Kingma, D. P. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021. 3

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS. 3

Tian, Y., Fan, L., Isola, P., Chang, H., and Krishnan, D. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48382–48402. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/971f1e59cd956cc094da4e2f78c6ea7c-Paper-Conference.pdf. 2, 6

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017. 2

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021. 5

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 2, 6

A Proofs

Proof of Proposition 2.1. **Qualitative clause.** Disintegrate S as $S(dx) = S_z(dx)\,\nu(dz)$. By hypothesis there exist measurable fibre selections $x^-(z), x^+(z) \in \operatorname{supp} S_z$ whose induced energy laws under ν differ; measurable selection is standard under regularity of S_z (Castaing & Valadier, 1977). Define

$$Q^{\pm}(dx) \;=\; \int \delta_{x^{\pm}(z)}\,\nu(dz).$$

Since $\phi(x^{\pm}(z)) = z$ for every z, we have $\phi_{\#}Q^{\pm} = \nu$, so $Q^{\pm} \in \Pi_\phi(\mu, \Sigma)$ and $\mathrm{FID}_\phi(R, Q^-) = \mathrm{FID}_\phi(R, Q^+)$. By construction $(E_P)_{\#}Q^{\pm}$ is the law of $E_P(x^{\pm}(\cdot))$ under ν, and these laws differ by hypothesis, giving $(E_P)_{\#}Q^- \neq (E_P)_{\#}Q^+$.

Quantitative clause. The hypothesis specifies that Z, a, Δ, and selections $x^-(z), x^+(z)$ exist with $E_P(x^-(z)) \le a$ and $E_P(x^+(z)) \ge a + \Delta$ for $z \in Z$, and $x^-(z) = x^+(z)$ for $z \notin Z$. Then Q^- and Q^+ agree off Z, and the Z-indexed component of Q^- assigns energy at most a while that of Q^+ assigns energy at least $a + \Delta$. The energy pushforwards therefore satisfy

$$(E_P)_{\#}Q^-\big((-\infty, a]\big) - (E_P)_{\#}Q^+\big((-\infty, a]\big) \;\ge\; \alpha,$$

with the symmetric inequality on $[a + \Delta, \infty)$. Any transport plan between $(E_P)_{\#}Q^-$ and $(E_P)_{\#}Q^+$ must therefore move at least α mass across the interval $[a, a + \Delta]$, which has length Δ. The one-dimensional Wasserstein distance consequently satisfies

$$W_1\big((E_P)_{\#}Q^-, (E_P)_{\#}Q^+\big) \;\ge\; \alpha\Delta. \quad \square$$

$\square$

B Experiments

Compute details. All density scoring, metric computation, and downstream training jobs were run on NVIDIA V100 GPUs. Energy scores are computed once per dataset and reused across all ELD analyses; after scoring, ELD itself is computed from sorted one-dimensional energy values and is negligible compared with image scoring or representation training.

B.1 Diagnostics

.

Regime	# datasets	Mean shift	Mean spread	Mean pELD
Over-concentrated	3	-4.43	0.52	.281
Broad / shifted	7	-5.41	1.47	.284
Noise-like / high-energy	4	2.92	0.22	.300

Table 3: **Coarse regime diagnostics.** Mean shift and spread separate the main regimes; pELD captures distributional mismatch beyond these summaries.

Simple diagnostics give a coarse regime map, not a substitute for ELD. Mean energy shift and spread ratio provide a useful low-dimensional summary of how a source departs from ImageNet: they separate compression, displacement, and high-energy mismatch. Table 3 shows that these summaries recover the coarse regimes used in the main text. However, they do not determine the full energy distribution. Sources within the same regime can have different tail structure and different pELD, so the Wasserstein comparison remains necessary for quantifying residual distributional mismatch.

Generic feature-density energies are not substitutes for ELD. We also test whether simpler surrogate energy models reproduce the ranking induced by the learned ImageNet density. Specifically, we fit three lightweight alternatives on ImageNet: a diagonal Gaussian over frozen ResNet-18 features, a diagonal Gaussian over low-frequency greyscale DCT coefficients, and a Gaussian mixture model over PCA-compressed DCT features. These alternatives capture feature novelty or low-level image statistics, but they are unable to recover the ELD ranking from the primary density model. Their rank

Surrogate energy model	Description	Spearman ρ with pELD
ResNet-feature Gaussian	Diagonal Gaussian on frozen features	-0.22
DCT Gaussian	Diagonal Gaussian on low-frequency DCT coefficients	-0.28
DCT-GMM	GMM on PCA-compressed DCT coefficients	-0.26

Table 4: **Simple surrogate energies do not reproduce ELD.** Feature- and DCT-based density surrogates fail to recover the ranking induced by the learned ImageNet density model, indicating that ELD is not reducible to generic feature novelty or low-level image-statistics mismatch.

correlations with primary pELD are negative or near zero: $\rho = -0.22$ for ResNet-feature Gaussian energy, $\rho = -0.28$ for diagonal DCT-Gaussian energy, and $\rho = -0.26$ for DCT-GMM energy. This suggests that ELD is not merely measuring generic feature outlierness, spectrum mismatch, or distance from simple image statistics. The choice of p_P matters: ELD is intended to compare datasets under a real-image density model, not under arbitrary feature-space novelty scores.

B.2 Synthetic Source Visualizations

Figures 6 and 7 show sampled images from the synthetic sources used in the DINO/ViT-S and Learning-to-See experiments. The visual differences across sources are substantial, but visual appearance alone does not determine the energy profile: visually diverse generators can still be shifted or compressed along the real-image likelihood axis, while simpler sources can occupy distinct energy regimes. These grids provide qualitative context for the quantitative energy audit in the main text.

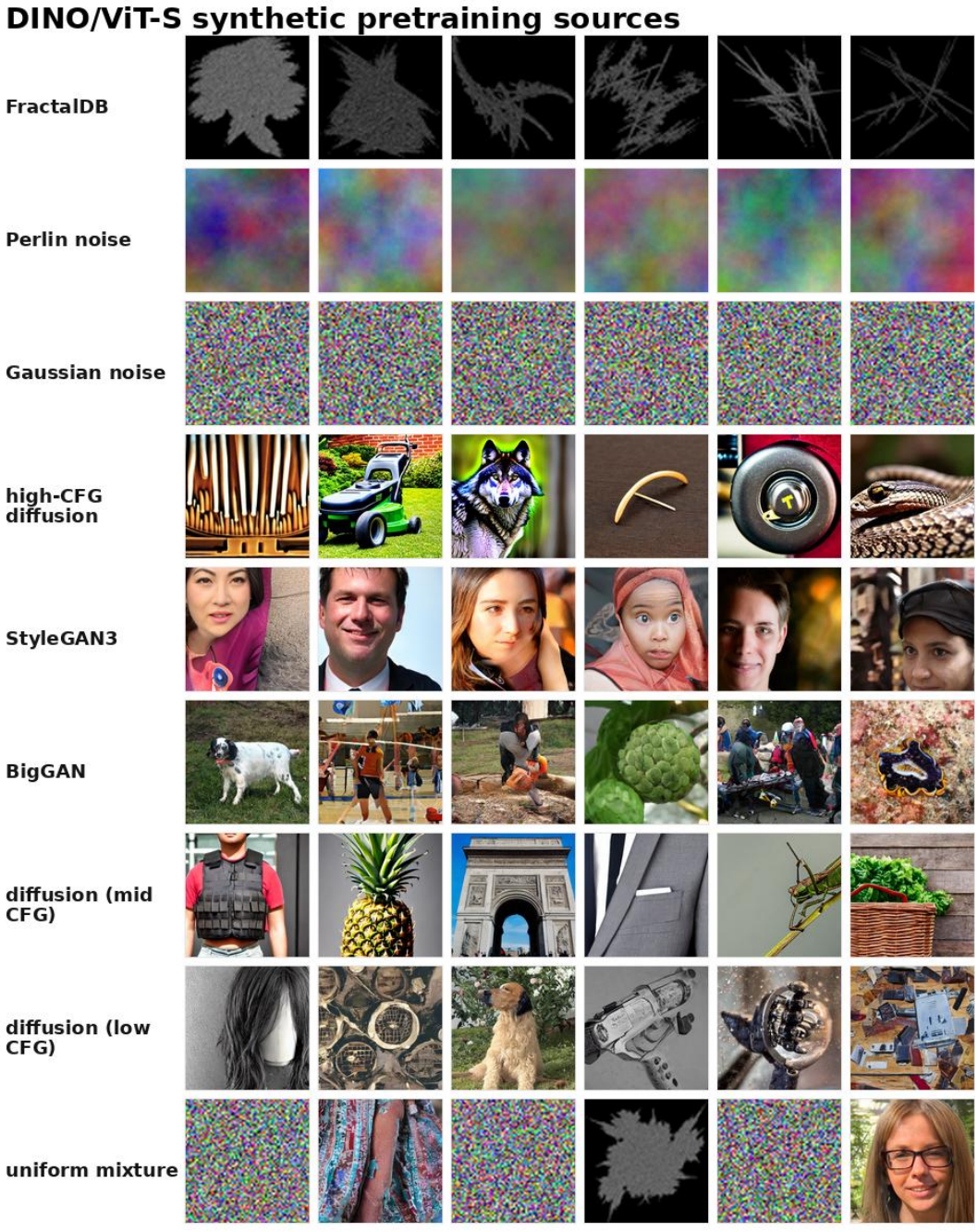

Figure 6: **DINO/ViT-S synthetic pretraining sources.** This grid highlights the range from noise and procedural structure to generator samples; ELD measures how these sources populate ImageNet likelihood levels.

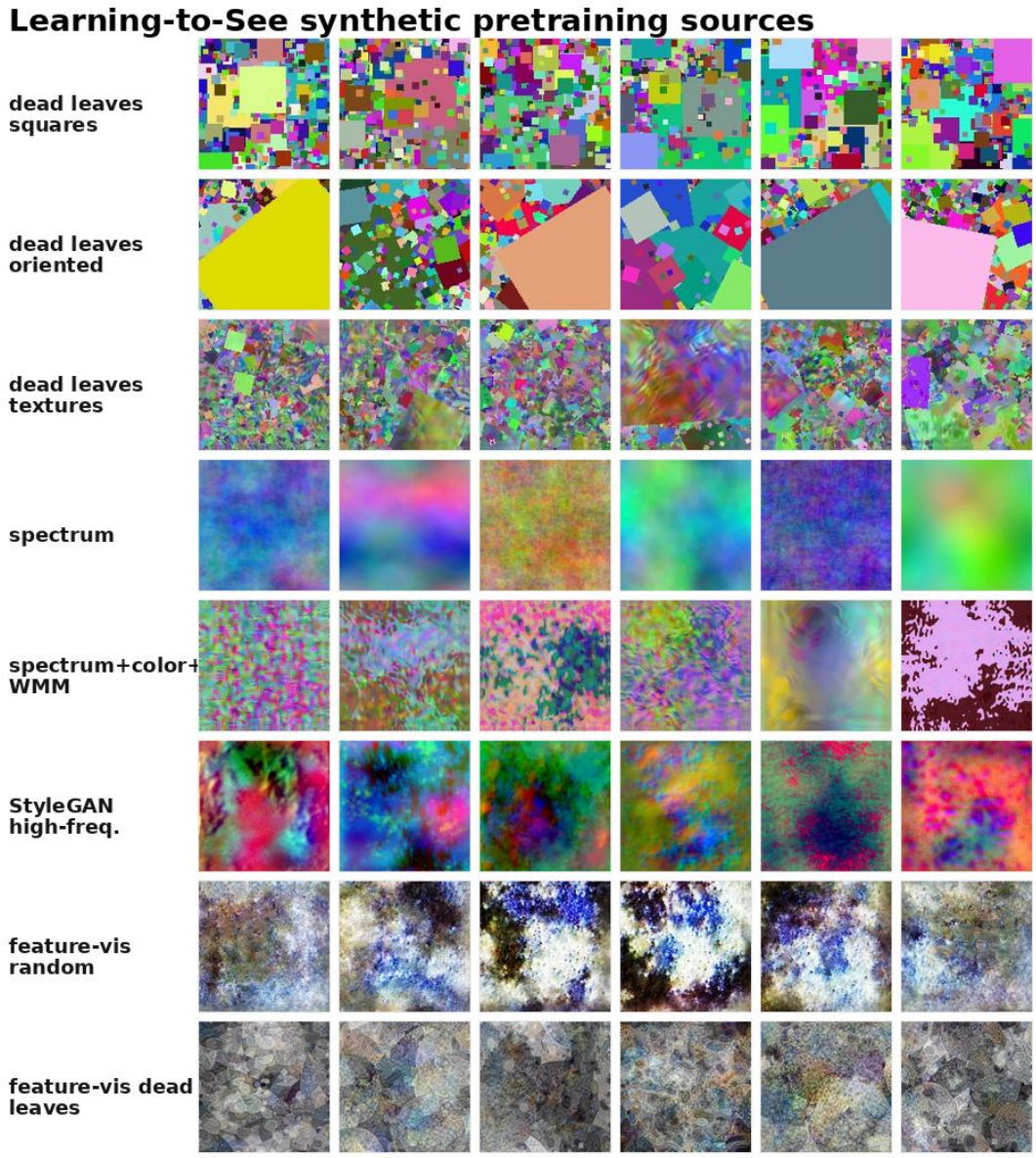

Figure 7: **Learning-to-See synthetic pretraining sources.** This grid visualises the range from dead-leaves and statistical textures to untrained StyleGAN and feature-visualization samples; ELD measures how these families populate ImageNet likelihood levels.