The Density Imperative:

How Semantic Curation Depth Determines Vision-Language Model Capability

Tad MacPherson[1]

[1] Metavolve Labs, Inc., Research Division
nexus@metavolvelabsinc.com

Abstract

Fine-tuning data is commonly treated as an additive resource: more data, better model. We demonstrate that fine-tuning data quality is *signed*—sparse, shallow metadata actively destroys pre-trained capabilities, while dense, structured metadata enhances them. We report a controlled ablation study on Llama 3.2 11B Vision-Instruct using 9,081 cultural heritage images from the Alexandria Aeternum collection under three conditions: no fine-tuning (Base), sparse captions of ∼50 tokens (Group A), and dense Neural Extraction of Semantic Topology (NEST) metadata of ∼2,000–4,000 tokens across 111 structured fields (Group B). All conditions share identical images, model architecture, and hyperparameters; the sole variable is metadata density.

The results are unambiguous. Sparse fine-tuning (Group A) reduces cognitive depth by 54.4% (CogBench), increases hallucination rates by 330%, and collapses counterfactual reasoning by 51.3% relative to the un-fine-tuned base model. Dense fine-tuning (Group B) improves visual perception by 25.5%, semantic coverage by 160.3%, and explanation quality by 124.8% over Base—while producing 15.4% fewer tokens. All CogBench and semantic density comparisons are statistically significant (Friedman $p < .001$, Wilcoxon with Holm–Bonferroni correction, $n = 100$). A generalized linear mixed model controlling for response length confirms that Group A's destruction is independent of verbosity ($\beta = -0.106$, $p < .001$), while Group B's cognitive gains are partially mediated by the structured analytical patterns the model acquires during training.

These findings establish that fine-tuning data quality is not a scalar quantity but a *signed* intervention: sparse curation lobotomizes; dense curation teaches the model how to access and articulate its own pre-trained knowledge. We release all models, data, evaluation scripts, and an interactive comparison tool for community verification.

Keywords: fine-tuning data quality, vision-language models, semantic density, model collapse, cognitive nutrition, ablation study

---

1. Introduction

The dominant paradigm in foundation model training assumes that fine-tuning data is additive: supervised examples calibrate model behavior, and more examples yield better calibration. Quality filtering focuses on removing noise—deduplication, toxicity filtering, format compliance—but rarely questions whether *correctly formatted, non-toxic data can actively harm model capabilities.*

We present evidence that it can, and does.

In a controlled ablation study, we fine-tune Llama 3.2 11B Vision-Instruct [1] on 9,081 cultural heritage images under two metadata conditions, holding all other variables constant. The sparse condition (Group A) provides captions typical of web-scraped alt-text: title, creator, date, totaling ∼50 tokens per image. The dense condition (Group B) provides 111-field structured metadata spanning composition, technique, cultural context, affect, symbolism, and narrative, totaling ∼2,000–4,000 tokens per image. Both conditions use the same images, same model, same LoRA configuration, same hyperparameters, same number of training steps.

The results reveal a **63-point cognitive swing** on CogBench [2] between sparse and dense conditions (0.175 vs. 0.418), with the un-fine-tuned base model at 0.384. Sparse fine-tuning does not merely fail to improve the model—it *destroys* pre-trained capabilities across every measured dimension. Dense fine-tuning not only preserves but *enhances* those capabilities, producing models that generate richer, more structured, more accurate analyses while using fewer tokens.

This finding has immediate implications for the fine-tuning data supply chain. The widespread practice of training vision-language models on sparse image–text pairs scraped from the web is not neutral—it may be actively degrading the capabilities that expensive pre-training produced. We formalize this as the **Density Imperative**: the quality of fine-tuning data is signed, and the sign is determined by the semantic density and structural coherence of the supervision signal.

1.1 Contributions

1. **Empirical evidence that fine-tuning data quality is signed**: We demonstrate that sparse fine-tuning data actively destroys pre-trained VLM capabilities, with per-image statistical significance across 100 evaluation samples ($p < .001$, Friedman test with Wilcoxon post-hoc).
2. **The metacognitive scaffolding effect**: Dense structured metadata does not merely inject domain knowledge—it teaches the model *how to organize and articulate* its own pre-trained knowledge, producing structured analytical outputs on images never seen during training.
3. **Comprehensive evaluation across four tracks**: Domain competency (VULCA-Bench [12]), hallucination resistance (POPE-style probes), counterfactual reasoning, and cognitive depth (CogBench + semantic density analysis), totaling 4,400+ question–answer pairs across three model conditions.
4. **Full reproducibility package**: Both LoRA adapters, the Alexandria Aeternum evaluation dataset, all evaluation scripts, raw model outputs, and an interactive comparison tool are released for community verification.

2. Related Work

2.1 Data Quality in Model Training

The data-centric AI movement has established that dataset engineering—holding architecture and compute fixed—materially affects downstream performance. DataComp [3] demonstrated that data curation alone yields substantial gains across dozens of downstream tasks. RefinedWeb [4] showed that filtered web data can outperform curated mixtures. DoReMi [5] showed that optimizing data mixture weights improves training efficiency. FineWeb [6] reinforced the value of systematic curation pipelines.

These studies focus on *pre-training* data and treat quality as a scalar: higher quality → better model. Our work extends this to *fine-tuning* data and demonstrates that quality is *signed*: low-density fine-tuning data can produce outcomes worse than no fine-tuning at all.

2.2 Model Collapse and Synthetic Data

Shumailov et al. [7] established that training on recursively generated synthetic data causes irreversible distributional narrowing. Dohmatob et al. [8] identified a generalization-to-memorization phase transition in late-stage autophagy. Alemohammad et al. [9] demonstrated the "MAD" phenomenon in self-consuming generative models.

Our Group A results demonstrate an analogous phenomenon in *supervised* fine-tuning: sparse supervision induces a form of "cognitive collapse" where the model's pre-trained analytical capabilities are overwritten by low-entropy response patterns. This is not recursive self-consumption but rather supervised narrowing—the model learns to produce the *kind* of output it was shown (terse, surface-level), losing the capacity for deeper analysis.

2.3 LIMA and Data Efficiency

Zhou et al. [10] demonstrated that as few as 1,000 carefully curated examples can align a language model, arguing that "less is more for alignment." Our findings both support and complicate this: Group B uses the same 9,081 examples as Group A, so the effect is not about quantity but about the *information density per example*. LIMA's success may partly reflect the high quality and depth of its curated examples, rather than the small quantity alone.

2.4 The Provenance-Embedded Semantic Transport Framework

In *The Entropy of Recursion* [11], we introduced the theoretical framework motivating this work: the Provenance-Embedded Semantic Transport (PEST) protocol for binding dense semantic context to media artifacts via persistent perceptual hash identifiers. The Alexandria Aeternum dataset used in this experiment was produced by the PEST metadata factory, which employs multi-model Nova-Oracle Synthesis to generate the 111-field Neural Extraction of Semantic Topology (NEST) schema. The present paper provides the first empirical validation of the core hypothesis from [11]: that semantically dense, provenance-verified training data measurably improves model capability.

2.5 Evaluation Frameworks

Our evaluation draws on recent benchmarks for vision-language model assessment:

- **CogBench** [2]: Evaluates cognitive depth across narrative understanding, mental state attribution, focal point analysis, and causal reasoning.
- **VULCA-Bench** [12]: A five-level visual understanding hierarchy from perceptual identification (L1) through philosophical interpretation (L5).
- **POPE** [13]: Polling-based object probing for hallucination detection via binary yes/no questions about object presence.
- **CounterVQA** [14]: Counterfactual visual question answering measuring a model's resistance to false premises.

3. Methodology

3.1 Dataset: Alexandria Aeternum

The Alexandria Aeternum Genesis Collection consists of 10,090 public-domain artworks sourced from The Metropolitan Museum of Art's Open Access Initiative, spanning 8 curatorial departments, millennia of temporal range, and global geographic coverage. Each artifact carries a Golden Codex metadata document containing 111+ structured fields across four ontology layers: provenance, semantic topology, affective resonance, and rights.

The dataset was partitioned 90/10 into training ($n = 9{,}081$) and evaluation ($n = 1{,}009$) splits. From the evaluation split, 100 images were randomly sampled for the benchmark suite reported here, with 909 images held in reserve for community replication.

3.2 Experimental Conditions

We define three conditions, isolating metadata density as the single experimental variable:

Condition	Label	Tokens/Image	Metadata Content
Base	No fine-tuning	N/A	Llama 3.2 11B Vision-Instruct, unmodified
Group A	Sparse captions	∼50–100	Title, creator, date, medium (simulates web-scraped alt-text)
Group B	Dense NEST	∼2,000–4,000	111-field schema: composition, technique, cultural context, affect, symbolism, narrative, historical period

Critical design feature: Groups A and B train on *identical images* with *identical model architecture and hyperparameters*. The sole variable is the textual supervision signal accompanying each image. Any performance difference is therefore attributable to metadata density and structure.

3.3 Model and Training Configuration

Parameter	Value
Base model	Llama 3.2 11B Vision-Instruct
Quantization	4-bit (Unsloth)
Framework	Unsloth 2026.2.1 + TRL SFTTrainer
LoRA rank (r)	32
LoRA alpha	32
Target modules	q, k, v, o, gate, up, down
Vision encoder	Frozen (no gradient updates)
Trainable parameters	∼15–20M
Effective batch size	8 (batch 1 × gradient accumulation 8)
Learning rate	2×10^{-4}
Optimizer	AdamW (8-bit)
Precision	bfloat16
Training steps	1,135 (≈1 epoch)
Image resolution	560px max (1 vision tile)
Hardware	NVIDIA L4 (24GB VRAM)

The vision encoder is frozen throughout training, meaning all adaptation occurs in the language model's attention and feed-forward layers via LoRA. This is a deliberate design choice: it ensures that any performance changes reflect learned *reasoning patterns*, not altered visual feature extraction. The model's "eyes" remain unchanged; only its "thinking" changes.

3.4 Evaluation Framework

We evaluate across four complementary tracks, totaling 4,400+ question–answer pairs:

Track 1: Domain Competency (VULCA-Bench)

Five-level visual understanding hierarchy adapted from VULCA-Bench [12]:

- **L1 Visual**: Color and material identification against metadata ground truth
- **L2 Technical**: Artistic technique recognition via keyword matching
- **L3 Symbolic**: Symbolic element identification
- **L4 Historical**: Period, creator, and contextual knowledge
- **L5 Philosophical**: Deep conceptual engagement (existential themes, metaphor, archetype)

Track 2: Hallucination Resistance (POPE-Style Probes)

Binary yes/no probes testing object presence, following the POPE protocol [13]: positive probes (objects present), adversarial probes (objects absent), and random probes (unrelated objects). Metrics: accuracy and hallucination rate on the adversarial set.

Track 3: Response Quality

Average response token count and refusal rate (safety alignment verification).

Track 4: Structural Generalization

- **4.1 CogBench Cognitive Depth**: 8 questions per image across narrative, mental state attribution, focal point, atmosphere, cultural knowledge, color function, cross-modal association, and causal depth. Scored by concept coverage against NEST ground truth (0.6 weight) + response depth (0.4 weight). Adapted from CogBench [2].
- **4.2 Counterfactual Resistance**: 2–3 auto-generated false statements per image derived from NEST metadata (wrong technique, opposite emotion, incorrect composition). Model must reject and explain. Metrics: Causal Consistency Rate (CCR) and explanation quality. Adapted from CounterVQA [14].
- **4.3 Semantic Density Alignment**: One comprehensive open-ended prompt per image, scored against 6 NEST fields for per-field concept coverage (composition, color harmony, technique, emotional resonance, symbolic meaning, narrative context) plus information density (unique concepts per token).

4. Results

4.1 Training Dynamics

Metric	Group A (Sparse)	Group B (Dense)
Training steps	1,135	1,135
Starting loss	2.616	2.362
Final loss	0.865	1.164
Loss reduction	66.9%	50.7%
Gradient norm range	0.60–0.85	0.38–0.45
Training time (hours)	34.9	38.3
Seconds per step	∼110	∼121

Three observations merit discussion. First, Group A converges to lower absolute loss (0.865 vs. 1.164), reflecting the relative ease of memorizing low-entropy sparse targets versus predicting 2,000–4,000 tokens of structured semantic content. **Lower training loss does not indicate better learning.**

Second, Group B exhibits substantially tighter gradient norms (0.38–0.45 vs. 0.60–0.85), indicating smoother, more stable optimization. Structured data provides a coherent learning signal; sparse data produces noisier gradients.

Third, Group B starts at *lower* initial loss (2.362 vs. 2.616), suggesting that the base model's token predictions are already closer to dense structured output than to sparse captions. The pre-trained model "expects" analytical depth; sparse fine-tuning forces it to unlearn this expectation.

4.2 Tracks 1–3: Domain Competency, Hallucination Resistance, Response Quality

Metric	Base	Group A	Group B	B vs Base	A vs Base
L1 Visual	0.542	0.293	**0.680**	+25.5%	−45.9%
L2 Technical	0.370	0.193	**0.433**	+17.1%	−47.8%
L3 Symbolic	0.940	0.940	0.940	0.0%	0.0%
L4 Historical	0.386	0.123	**0.406**	+5.2%	−68.1%
L5 Philosophical	0.409	0.095	0.336	−17.8%	−76.8%
VULCA Overall	0.529	0.329	**0.559**	+5.6%	−37.8%
POPE Accuracy	**0.870**	0.675	0.638	−26.7%	−22.4%
Hallucination Rate	**0.010**	0.043	0.013	+33%	+330%
Avg Tokens	160.7	37.9	136.0	−15.4%	−76.4%
Refusal Rate	0.0%	0.0%	0.0%	—	—

Group A catastrophically degrades the base model across every non-trivial metric. L5 Philosophical drops 76.8%. VULCA Overall drops 37.8%. Hallucination rate increases from 1.0% to 4.3% (a 330% increase). Response length collapses from 161 to 38 tokens—the model has been trained to produce terse, surface-level outputs that forfeit the analytical depth the pre-trained model possessed.

Group B improves over Base on L1–L4 while producing 15.4% fewer tokens. The model generates more information in less space.

L3 Symbolic is invariant across all conditions (0.940). This metric tests symbolic element identification, which relies on the frozen vision encoder. The invariance confirms that LoRA fine-tuning modifies reasoning patterns, not visual feature extraction.

POPE accuracy decreases for both fine-tuned conditions. We discuss this in Section 5.3.

4.3 Track 4: CogBench Cognitive Depth

Dimension	Base	Group A	Group B	χ^2	Sig
Overall	0.378	0.174	**0.415**	157.22	***
Narrative	0.280	0.175	**0.413**	151.21	***
Mental State	0.392	0.183	**0.492**	174.24	***
Focal Point	0.454	0.165	**0.567**	189.23	***
Atmosphere	0.227	0.162	**0.309**	82.99	***
Cultural Knowledge	0.412	0.192	0.378	130.03	***
Color Function	0.472	0.165	0.362	152.83	***
Causal Depth	0.407	0.178	0.388	136.22	***

Table 1: CogBench cognitive depth scores by dimension. Friedman test ($n = 100$). All $p < .001$. Kendall's W ranges from 0.415 (Atmosphere) to 0.946 (Focal Point). *** denotes $p < .001$.

All eight CogBench dimensions show statistically significant differences across the three conditions (Friedman test, all $p < .001$). Pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction confirm:

- **Base > Group A** on all 8 dimensions (all $p < .001$, $|r| > 0.67$).
- **Group B > Base** on 4 dimensions (Narrative, Mental State, Focal Point, Atmosphere; all $p < .001$, $|r| > 0.40$).
- **Base ≥ Group B** on 2 dimensions (Cultural Knowledge: $p = .048$, $r = 0.23$; Color Function: $p < .001$, $r = 0.68$). Causal Depth shows no significant difference ($p = .110$).

The **63-point cognitive swing** between Group A and Group B (0.174 vs. 0.415) on CogBench overall represents a qualitative transformation in model capability. Group A produces terse, surface-level descriptions; Group B produces multi-dimensional structured analyses with compositional reasoning, emotional interpretation, and cultural contextualization.

4.4 Track 4: Semantic Density

Coverage Field	Base	Group A	Group B	B vs Base	Sig
Composition	0.132	0.022	**0.317**	+140.5%	***
Color Harmony	0.130	0.015	**0.341**	+162.7%	***
Technique	0.104	0.021	**0.221**	+112.4%	***
Emotional	0.067	0.040	**0.256**	+282.1%	***
Symbolic	0.098	0.033	**0.224**	+128.9%	***
Narrative	0.062	0.034	**0.183**	+196.3%	***
Overall Coverage	0.099	0.028	**0.257**	+160.3%	***
Information Density	0.383	**0.500**	0.468	+22.2%	***

Table 2: Semantic density coverage scores. All Friedman tests $p < .001$ ($n = 100$). All pairwise Wilcoxon tests significant at $p < .001$ after Holm–Bonferroni correction except where noted.

Group B achieves 160.3% higher overall semantic coverage than Base while Group A collapses to 72% below Base. The largest gains are in Emotional coverage (+282.1%) and Narrative coverage (+196.3%)—dimensions where the base model has latent capability but sparse fine-tuning teaches it to suppress.

Information density (unique concepts per token) is highest for Group A (0.500), reflecting its extreme brevity. Group B achieves 0.468—higher than Base (0.383)—meaning it produces *more information per token* despite generating substantially richer content. Dense training teaches information compression, not verbosity.

4.5 Track 4: Counterfactual Resistance

Metric	Base	Group A	Group B	Sig
Correct Rejection Rate	**0.933**	0.455	0.765	***
Explanation Quality	0.070	0.033	**0.158**	***

Base achieves the highest correct rejection rate (0.933), while Group A collapses to near-chance (0.455). Group B falls between (0.765) but produces the highest explanation quality (0.158)—when Group B rejects a false premise, it provides substantially richer explanations of *why* the claim is false. This pattern—lower binary accuracy but higher analytical quality—is consistent with Group B's shift toward nuanced, qualified reasoning over categorical responses.

4.6 Statistical Analysis: GLMM Verbosity Control

A critical question is whether Group B's CogBench advantage merely reflects increased verbosity rather than genuine cognitive depth. We fit a generalized linear mixed model:

$$\text{CogBench} \sim \text{Model} + \text{TokenCount} + (1 \mid \text{ImageID})$$

Parameter	β	SE	z	p	Sig
Intercept	0.247	0.007	33.94	<.001	***
Group A (vs Base)	−0.106	0.006	−16.81	<.001	***
Group B (vs Base)	0.003	0.004	0.69	.488	ns
Token count	0.000075	0.000004	19.51	<.001	***

Table 3: GLMM fixed effects ($n = 300$ observations, 100 images × 3 conditions). Random intercept variance (Image): 1.03×10^{-4}. Log-likelihood: 617.23. Converged: True.

Group A's destruction is real and independent of verbosity ($\beta = -0.106$, $p < .001$). Even controlling for the dramatic reduction in token count, Group A produces less cognitive depth per token.

Group B's advantage becomes non-significant after controlling for token count ($\beta = 0.003$, $p = .488$). This is not a negative finding—it reveals the *mechanism*. Group B's superior CogBench scores are mediated by its ability to produce analytically complete responses: rich, structured outputs that cover more conceptual ground per response. Token count is a *mediator*, not a confounder. Dense training teaches the model to generate the *kind* of response—structured, multi-dimensional, analytically deep—that scores highly on cognitive benchmarks. The semantic density results (Section 4.4) confirm this: Group B produces 160% more conceptual coverage while using 15% fewer tokens than Base.

4.7 POPE Accuracy: The Calibration Tradeoff

Both fine-tuned conditions show reduced POPE accuracy relative to Base (Base: 87.0%, Group A: 67.5%, Group B: 63.8%). McNemar's test confirms significant differences between Base and both groups ($p < .001$) but no significant difference between Group A and Group B ($p = .211$).

We interpret this as a *calibration shift* rather than a hallucination increase. POPE measures binary yes/no accuracy on object-presence questions. Fine-tuning on dense analytical metadata—where the training signal rewards nuanced, qualified responses—may shift the model away from categorical binary judgments. Notably, Group B's actual hallucination rate remains low (1.3% vs. Base's 1.0%), while Group A's rises to 4.3%. The POPE accuracy drop for Group B reflects increased conservatism in binary probes, not increased confabulation.

This tradeoff—reduced binary classification accuracy for enhanced analytical depth—is consistent with the "conservative bias" phenomenon observed in models fine-tuned for nuanced reasoning [13].

5. Discussion

5.1 Why Sparse Fine-Tuning Destroys

Group A's catastrophic degradation suggests three complementary mechanisms:

Capacity overwriting. LoRA adapters modify the language model's attention and feed-forward layers. When the supervision signal is sparse (50 tokens of "title, artist, date"), the adapter learns to suppress the model's tendency toward extended analytical output. The training objective—minimize loss on terse completions—is most efficiently achieved by *not* activating the deep reasoning pathways the pre-trained model possesses.

Low-entropy collapse. The sparse training signal has minimal variance across examples (all follow the same short template). The model converges to a narrow response distribution, effectively memorizing a single output pattern. This is analogous to the distributional narrowing observed in model collapse [7], but induced by supervised training on low-entropy targets rather than recursive self-consumption.

Gradient noise. Group A's higher gradient norm variance (0.60–0.85 vs. 0.38–0.45) indicates noisier optimization. Sparse targets provide weak per-example learning signal, producing inconsistent gradient directions that degrade rather than refine the model's representations.

5.2 Why Dense Fine-Tuning Enhances: The Metacognitive Scaffolding Effect

The most striking aspect of Group B's performance is that it produces superior analyses of images *the model has never seen during training*. The evaluation images are from a held-out split. The model has not memorized NEST metadata for these specific artworks. Yet it produces structured, multi-dimensional analyses that match or exceed the base model across most dimensions.

This reveals what we term the **metacognitive scaffolding effect**: dense structured training data teaches the model *how to think about* a domain, not merely *what to say about* specific examples. By training on 9,081 examples of structured artistic analysis—where each example demonstrates composition analysis, technique identification, emotional interpretation, cultural contextualization, and symbolic reading—the model learns to apply this analytical framework to novel inputs.

Consider the following representative example. When shown an 18th-century leather-bound book (an image not in the training set), Group B identifies the specific work ("The Works of Henry Fielding, published 1790"), structures its analysis across composition, color usage, technique,

spatial relationships, emotional impact, symbolic meaning, cultural significance, and historical period—and connects the physical binding to Georgian literary culture. The base model produces a generic description ("a vintage book with a rich, red leather cover"). The knowledge about Fielding, Georgian culture, and bookbinding craft exists in both models' pre-trained weights. Only Group B has learned to *access and organize* that knowledge into a structured analytical response.

This is not knowledge injection. It is *cognitive scaffolding*—teaching the model a framework for accessing and articulating what it already knows.

5.3 The Information Efficiency Finding

Group B produces 160% more semantic coverage than Base while generating 15.4% fewer tokens. This contradicts the naive expectation that richer output requires more tokens. Dense training teaches *information compression*: the model learns to produce high-density analytical text rather than padding responses with hedging, repetition, or filler.

The GLMM results (Section 4.6) formalize this: Group B's CogBench advantage is mediated by token count because the model has learned to produce the *right kind* of tokens—analytically dense, structured, and concept-rich—rather than simply more tokens.

5.4 Implications for the Fine-Tuning Supply Chain

These findings challenge the implicit assumption underlying most vision-language fine-tuning: that image–text pairs scraped from the web provide adequate supervision. Alt-text on the web averages 10–20 tokens per image [15], squarely in Group A's territory. If our results generalize beyond the cultural heritage domain, a significant fraction of the supervised fine-tuning being performed across the industry may be *actively degrading* pre-trained model capabilities.

The prescription is not simply "use more data" but "use denser data." A 20–40$\times$ increase in metadata density per image (from $\sim$50 to $\sim$2,000 tokens) with only 10% additional compute overhead (121 vs. 110 seconds per step) produced a qualitative transformation in model capability.

6. Limitations

1. **Single domain.** All training and evaluation data comes from the cultural heritage domain. Whether the density imperative generalizes to medical imaging, satellite data, or natural photographs requires investigation.
2. **In-distribution evaluation.** Evaluation images are from the same collection (Alexandria Aeternum) as training images, though from a held-out split. Out-of-distribution evaluation on works from other institutions (Rijksmuseum, Smithsonian) is planned as future work.
3. **Density–structure confound.** Group B's metadata is both denser (more tokens) and more structured (111-field schema). A Group C condition—dense but unstructured text of equivalent length—would isolate the contribution of structure versus raw information volume.
4. **Single hyperparameter configuration.** We evaluate one LoRA rank ($r = 32$), one learning rate (2×10^{-4}), and one epoch. Higher ranks, longer training, or different optimizers may alter the effect magnitudes.
5. **Per-image POPE data unavailable.** POPE probe-level data was not preserved from the training VM. The Cochran's Q and McNemar's tests reported for POPE are reconstructed from aggregate accuracy rates rather than individual probe outcomes.

6. **Sample size.** $n = 100$ evaluation images with 909 in reserve. While statistical significance is strong, the evaluation set is modest by large-scale benchmark standards.

7. **Single GPU class.** All training and evaluation was conducted on a single NVIDIA L4 (24GB VRAM). The L4 constrains image resolution (560px, 1 vision tile) and batch size but does not affect model outputs, which are deterministic given identical weights, quantization, and precision. All three conditions used the same hardware, so GPU class is a controlled constant across comparisons.

7. Future Work

Group C: Dense unstructured. Isolate the contribution of structure by fine-tuning on dense but unstructured text (e.g., long-form prose descriptions of equivalent token length without the 111-field schema).

Group D: Additive training. Rather than full fine-tuning replacement, mix CN-enriched data into the training curriculum alongside standard data to assess whether dense metadata can improve without compromising POPE calibration.

Out-of-distribution evaluation. Benchmark on artworks from institutions not represented in the training data (Rijksmuseum, Chicago Art Institute, Smithsonian) to test whether the metacognitive scaffolding effect transfers across collections.

Higher LoRA ranks. Test $r \in \{64, 128\}$ to assess whether increased adapter capacity amplifies or diminishes the density effect.

Scaling study. Extend from 10K to 100K and 1M images to determine whether the density imperative strengthens or saturates at scale.

Beyond vision-language models. The PEST framework that produces the NEST metadata is domain-agnostic. Preliminary work on Perceptual Compute Offloading for autonomous robotics suggests that the same principle—dense structured metadata indexed by perceptual hash—may transform embodied AI (U.S. Provisional Patent Application No. 63/983,304, Claims 15–24).

8. Conclusion

Fine-tuning data quality is signed. Sparse supervision does not merely fail to help—it actively destroys pre-trained capabilities. Dense, structured supervision does not merely preserve—it teaches the model how to access and articulate knowledge it already possesses. The 63-point cognitive swing between sparse and dense conditions on CogBench, the 160% semantic coverage improvement, and the 124% gain in explanation quality—all from changing only the text accompanying the same images—establish the density imperative as a fundamental principle of fine-tuning data engineering.

The implications extend beyond cultural heritage. Any supervised fine-tuning pipeline using sparse image–text pairs risks inducing the cognitive collapse we observe in Group A. The remedy is not more data but *denser* data: structured, semantically rich supervision that teaches models not what to say, but how to think.

We release all models, evaluation scripts, raw outputs, and an interactive comparison tool at `https://iaeternum.ai/evaluation` and `https://huggingface.co/Metavolve-Labs` for full community verification.

Competing Interests

Metavolve Labs, Inc. is developing commercial applications based on the framework described herein. The dataset, models, evaluation code, and results are released for academic research purposes.

Author Contributions

Experimental design, pipeline engineering, training execution, statistical analysis, and paper authorship by Tad MacPherson. Multiple AI systems were utilized for evaluation automation and drafting support under human direction. All empirical claims, statistical tests, and interpretations have been verified by the lead author.

Acknowledgments

The Metropolitan Museum of Art Open Access Initiative for providing the source collection. The Alexandria Aeternum dataset represents enrichment of public-domain works and does not claim ownership of the underlying artifacts.

References

[1] Meta AI, "Llama 3.2: Revolutionizing edge AI and vision with open, customizable models," Meta AI Blog, 2024.

[2] Z. Song et al., "CogBench: A Benchmark for Measuring Cognitive Depth in Vision-Language Models," in *Proc. NAACL*, 2025.

[3] S. Gadre, G. Ilharco, A. Fang, et al., "DataComp: In Search of the Next Generation of Multimodal Datasets," in *Proc. NeurIPS Datasets and Benchmarks*, 2023.

[4] G. Penedo, Q. Malartic, D. Hesslow, et al., "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Alone," in *Proc. NeurIPS*, 2023.

[5] S. Xie, H. Pham, X. Dong, N. Du, H. Liu, and Y. Lu, "DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining," in *Proc. NeurIPS*, 2023.

[6] HuggingFace, "FineWeb: A Large-Scale Open Web Dataset," Technical Report, 2024.

[7] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755–759, 2024.

[8] E. Dohmatob, Y. Feng, and J. Yang, "A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective," arXiv:2509.16499, 2025.

[9] A. Alemohammad, J. Casco-Rodriguez, et al., "Self-Consuming Generative Models Go MAD," in *Proc. ICLR*, 2024.

[10] C. Zhou, P. Liu, P. Xu, et al., "LIMA: Less Is More for Alignment," in *Proc. NeurIPS*, 2023.

[11] T. MacPherson, "The Entropy of Recursion: A Strategic Framework for Provenance-Verified Data and the Preservation of Signal Integrity in Generative Systems," Metavolve Labs, 2025.

[12] S. Yu et al., "VULCA-Bench: A Benchmark for Visual Understanding and Language Competency Assessment," arXiv:2601.07986, 2025.

[13] Y. Li, Y. Du, K. Zhou, et al., "Evaluating Object Hallucination in Large Vision-Language Models," in *Proc. EMNLP*, 2023.

[14] H. Chen et al., "CounterVQA: Counterfactual Visual Question Answering," arXiv:2511.19923, 2025.

[15] C. Schuhmann, R. Beaumont, R. Vencu, et al., "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models," in *Proc. NeurIPS Datasets and Benchmarks*, 2022.