Diversity Scaling Laws for Embodied Imitation: Unseen Success Scales with Coverage Radius, Not Episode Count

Liz Lemma Future Detective

January 17, 2026

Abstract

Recent meta-analysis of robotics scaling laws finds that downstream success improves as a power law in data and model size, but that scaling on unseen tasks is markedly weaker than on seen tasks. We formalize this gap by treating embodied datasets as sets of contexts (scene-object-goal configurations) and defining diversity via geometric coverage. Our first contribution is a clean problem formulation for unseen-task scaling in imitation learning: given a dataset of episodes collected from contexts $\mathcal{S}$, characterize the best-achievable success on test contexts in terms of the dataset's coverage radius $r(\mathcal{S})$. Under a mild Lipschitz transfer assumption in context space, we prove matching upper and lower bounds showing that unseen-task error is $\Theta(r(\mathcal{S}))$ (up to system error). We then relate $r(\mathcal{S})$ to the number of distinct contexts and the intrinsic dimension of the context distribution, yielding a power law exponent $\alpha_{\text{unseen}} = -1/d_{\mathcal{X}}$ (up to logarithmic factors). Our second contribution is an operational estimator of effective diversity $D_{\text{eff}} = r^{-d_{\mathcal{X}}}$ and a dataset construction procedure (farthest-first context selection) that increases D_{eff} faster than naive data collection. Finally, we propose experiments in procedurally generated embodied suites where we independently vary episode count and context coverage; we predict that unseen success curves collapse when plotted against D_{eff}, turning the qualitative slogan "diversity matters" into a quantitative scaling law relevant for 2026-era robot data engines.

Table of Contents

1. 1. Introduction: unseen-task scaling gap in robotics; why episode count is insufficient; contributions and predictions.

2. 2. Related Work: neural scaling laws; imitation learning scaling; diversity/coverage in robotics datasets; emergence and generalization metrics.

3. 3. Problem Setup and Metrics: formalize tasks as context-indexed POMDPs; define seen vs unseen evaluation; define coverage radius and effective diversity D_{eff}; discuss when contexts are observed vs latent.

4. 4. Main Theoretical Results (Upper Bounds): Lipschitz transfer in context; show unseen error bounded by coverage radius; decomposition into (coverage + approximation/optimization + sim2real/system error).

5. 5. Main Theoretical Results (Lower Bounds): minimax lower bound via packing in context space; show any algorithm must incur error proportional to $r(\mathcal{S})$; discuss tightness and log factors.

6. 6. From Coverage to Power Laws: expected radius scaling under doubling dimension; derive $\alpha_{\text{unseen}} = -1/d_{\mathcal{X}}$; contrast with seen scaling; implications for dataset design.

7. 7. Algorithms: estimating diversity and constructing high-coverage datasets (farthest-first / k-center); computing D_{eff} from discrete factors or learned embeddings; complexity and approximation guarantees.

8. 8. Experimental Design (Recommended): procedural generators to vary count without diversity and diversity without count; evaluation protocol; predicted curve collapses; ablations across architectures (VLA vs VLM vs PVR).

9. 9. Discussion and Limitations: when Lipschitz transfer fails; representation learning issues; covariate shift beyond context; connections to compute-optimal scaling and inference-time compute.

10. 10. Conclusion: diversity scaling laws as a unifying explanation for weak unseen exponents; guidance for 2026 data engines and benchmarks.

1 Introduction

Empirical scaling laws have become a practical tool for anticipating performance gains from additional data and compute. In robotics, however, the most salient performance metric is rarely in-distribution return on the trajectories used for training; it is instead the ability to solve *unseen* tasks, i.e. tasks whose defining attributes (object instances, layouts, goals, dynamics parameters, sensing conditions) differ from those present in the training episodes. In this regime we repeatedly observe a gap between "seen-task" and "unseen-task" scaling: as the number of recorded episodes D increases, performance on previously encountered situations improves rapidly, while performance on held-out situations often improves slowly and saturates early. This discrepancy is not well explained by raw episode count alone.

The primary reason is that D conflates two distinct resources. On the one hand, additional episodes reduce statistical error on contexts already present in the dataset: repeated demonstrations can sharpen an estimate of the expert action distribution and reduce variance in behavior cloning. On the other hand, generalization to new tasks depends on *where* the data was collected: if most episodes arise from a small number of nearly identical contexts, then the learner has little basis for transferring to contexts far from those represented in the dataset. Consequently, two datasets with the same D may yield markedly different unseen-task performance, depending on their *coverage* of the task space.

We formalize this intuition by representing tasks as a family $\{\tau_x\}_{x\in\mathcal{X}}$ indexed by a context variable x taking values in a metric space $(\mathcal{X}, d)$. The context may be directly observed (e.g. a symbolic description of object types and goal positions), partially observed (e.g. an image), or latent and inferred via an embedding $\hat{\phi}(x)$. The metric d is chosen to reflect task similarity; it may be discrete (Hamming/edit distance over attributes), geometric (e.g. Wasserstein-like distances over layouts), or an embedding distance. A dataset induces a multiset of contexts $\mathcal{S} \subset \mathcal{X}$, and the relevant quantity for generalization is how well $\mathcal{S}$ approximates the test distribution over contexts. Intuitively, if every test context lies near some training context, then we can expect transfer; if there exist test contexts far from all training contexts, then no amount of repetition on the seen contexts can fix the resulting blind spots.

Our analysis proceeds under a transfer regularity assumption: success probability varies smoothly with context in the sense that small context perturbations cannot dramatically change success when we compare against a nearby expert. This hypothesis is deliberately weak: it does not require that the policy itself be Lipschitz in observations, nor that the transition dynamics vary smoothly in parameters; it asserts only that task performance does not change arbitrarily fast as a function of context. Under this assumption, unseen-task error can be controlled by the distance from an unseen

test context to its nearest context in $\mathcal{S}$. This yields a concrete operational recommendation: to improve generalization under a fixed episode budget D, we must preferentially allocate episodes to *new and distant* contexts rather than repeatedly sampling near-duplicates.

To make this principle quantitative, we introduce a coverage-based proxy for dataset "size" that we call *effective diversity*. Rather than measuring progress by D, we measure progress by how small a test context can be guaranteed to be from the training set in the worst case. This leads to a scalar summary of the dataset that increases when we add novel contexts and changes minimally when we repeat old ones. Under standard geometric regularity conditions on the context space, effective diversity admits an interpretable scaling with the number of distinct contexts K and yields a power-law prediction for unseen-task error whose exponent is governed by an intrinsic dimension $d_{\mathcal{X}}$ of the context distribution.

The resulting view clarifies several empirical phenomena. First, it predicts that unseen-task scaling exponents are typically smaller in magnitude than seen-task exponents: generalization is limited by how rapidly a finite set can cover a high-dimensional space, whereas interpolation among previously observed contexts can improve quickly with repeated samples. Second, it predicts that aggregating additional episodes without increasing coverage produces diminishing returns on unseen tasks, even if it continues to improve performance on the seen contexts. Third, it suggests that comparing learning systems by plotting error versus D can obscure the true driver of generalization; plotting error versus an estimate of effective diversity should produce a sharper, more stable relationship across architectures, laboratories, and collection protocols.

We also consider the dataset design problem that arises when a simulator or environment generator allows one to choose which contexts to sample. In that setting, the appropriate objective is explicitly geometric: select a set of contexts that minimizes the coverage radius over a prescribed candidate pool or evaluation set. This objective coincides with a classical k-center problem, and consequently admits efficient approximation algorithms with worst-case guarantees. The practical implication is that one can convert a fixed episode budget D into a principled context budget K (with $m = D/K$ episodes per context) and then choose contexts via a farthest-first traversal to maximize diversity.

Our contributions can be summarized as follows. (i) We propose a coverage-based effective diversity measure D_{eff} for robotics imitation datasets, along with a simple plug-in estimator based on nearest-neighbor distances over an evaluation pool. (ii) We provide a theoretical connection between coverage and unseen-task error under a Lipschitz transfer hypothesis, and we show that coverage is not merely sufficient but information-theoretically necessary for worst-case generalization. (iii) We derive a scaling prediction linking unseen-task error to effective diversity through an intrinsic dimension

parameter $d_{\mathcal{X}}$, yielding an exponent determined by geometry rather than by model class idiosyncrasies. (iv) We present a near-optimal context selection procedure, based on farthest-first k-center approximation, for maximizing effective diversity under a fixed episode budget.

The practical prediction is straightforward to falsify: when we evaluate on a preregistered set of unseen contexts, we should observe that performance as a function of D_{eff} is more stable than performance as a function of D, and that interventions which increase coverage (more distinct and more distant contexts) systematically outperform interventions which merely increase repetition. This perspective does not deny the importance of representation, optimization, or system imperfections; rather, it isolates a geometric bottleneck that persists even for an idealized learner and therefore must be addressed by data collection and benchmark design.

2 Related Work

Empirical *neural scaling laws* study how loss or error changes as a function of training resources such as dataset size, model capacity, and compute. In language modeling, a sequence of works established approximate power-law relationships between cross-entropy and the number of tokens and parameters, along with practical prescriptions for allocating compute between data and model size **??**. Related analyses appear in vision and multimodal learning, often emphasizing that the exponent can depend on the evaluation distribution and on whether one measures in-distribution likelihood versus downstream transfer **??**. Our setting differs in two respects: (i) the quantity of interest is *unseen-task* success in a family of interactive problems rather than predictive loss on held-out i.i.d. samples; and (ii) the relevant dataset resource is not solely the number of recorded episodes, but also the geometric *coverage* of the task space from which those episodes originate. Nevertheless, the methodological motivation is shared: we seek a low-dimensional summary of data that yields stable scaling behavior under controlled evaluation.

Scaling phenomena have also been investigated in reinforcement learning and decision making, where performance depends on both exploration and function approximation **??**. These works typically scale environment interactions or compute and evaluate on either the same task distribution or a suite of tasks. Our focus is narrower: we separate, by design, the statistical benefit of repeated episodes in the same context from the generalization benefit of collecting episodes in *new* contexts. This distinction is closely related to classical notions of distribution shift and coverage in offline RL, where lack of support of the behavior policy can lead to poor extrapolation **??**. However, rather than reasoning about state-action visitation coverage, we treat the task generator as inducing a metric space of *contexts* and ask

how far test tasks can be from the training contexts in that metric.

In imitation learning, sample complexity analyses typically quantify the number of expert demonstrations required to achieve low expected cost on a fixed task, under either i.i.d. supervised learning assumptions (behavior cloning) or interactive data collection (e.g. DAgger) **??**. Subsequent work refines these guarantees under partial observability, compounding error over horizon, and structured policy classes **??**. These results are primarily *within-task*: they hold when training and test rollouts are generated in the same environment instance. Our interest is complementary and orthogonal: even if we imagine an oracle imitation learner that matches the expert on each context represented in the dataset, performance can remain poor on *new* contexts due to the absence of any nearby training context to transfer from. In other words, our geometric bottleneck persists even when the usual supervised-learning estimation error is idealized away.

Several empirical robotics papers have emphasized that *diversity* of demonstrations and environments is a key driver of generalization. Large-scale robot learning efforts aggregate data across objects, scenes, and goals, and report improved robustness and transfer as the variety of training conditions increases **???**. Benchmarks for manipulation and navigation often operationalize generalization by holding out object instances, layouts, or goal specifications, sometimes reporting a steep gap between seen and unseen conditions **??**. A common limitation is that the data resource is reported as total transitions or episodes, which does not distinguish repetition from coverage. Our contribution is to propose an explicit metric-based notion of coverage radius and an associated effective diversity proxy that is intended to function analogously to "dataset size" in scaling plots, but tuned to unseen-task performance.

Metrics for dataset diversity in robotics have been proposed in several forms. At the trajectory level, one can measure variability via clustering in representation space, entropy of high-level labels, or diversity over goal specifications. At the state(-action) level, one can estimate coverage by visitation counts, kernel density estimates, or learned latent occupancy models, with ties to offline RL reliability **??**. These approaches are valuable but can be difficult to interpret across tasks and modalities, and they may conflate differences that are irrelevant for transfer with differences that are crucial. We instead posit that the environment generator naturally induces a context variable (observed or latent) and that an application-relevant metric d can be defined on contexts; our effective diversity is then a geometric quantity derived from nearest-neighbor distances in this metric space. This perspective aligns with work on *coresets* and dataset compression, where one chooses representative points to approximate a distribution or function class, often using k-center or farthest-first heuristics **??**. Our dataset design procedure is a direct instantiation of these ideas in the context space of tasks.

Generalization across tasks has also been studied through meta-learning

and representation learning, where a model is trained across a distribution of tasks and then adapted or evaluated on new tasks **??**. In such frameworks, one often introduces a latent task embedding or context encoder inferred from observations, which is then used to condition the policy **??**. These methods motivate our allowance for contexts that are not directly observed: in practice one may work with an estimated embedding $\hat{\phi}(x)$ and a corresponding metric in embedding space. Our theoretical development does not assume that the context is perfectly known; rather, it isolates the role of *geometric proximity* between training and test contexts, regardless of whether that proximity is computed in a symbolic space or a learned representation. Any embedding error is naturally absorbed into an irreducible system term in the performance bound.

Finally, the relationship between *intrinsic dimension* and rates of approximation has a long history in nonparametric statistics and learning theory. In metric spaces with finite doubling dimension, covering numbers control the number of points needed to approximate a distribution to a given resolution, yielding rates governed by the dimension rather than by ambient coordinates **?**. Our use of doubling dimension is in this tradition: it provides a principled way to translate "number of distinct contexts" into an expected coverage radius, and hence into a predicted rate for unseen-task error under Lipschitz transfer. In the next section we formalize the task family, the seen/unseen evaluation protocol, and the coverage-based quantities that instantiate this intuition.

3 Problem Setup and Metrics

We model a family of interactive tasks indexed by a *context* variable $x \in \mathcal{X}$, where $\mathcal{X}$ is equipped with a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. Intuitively, x encodes those aspects of an environment instance that are intended to vary across tasks (e.g., object identities and poses, scene layouts, goal specifications, or domain parameters). For each context x, we define a finite-horizon POMDP τ_x with horizon H. We do not require a particular parametric form of τ_x; it suffices that τ_x induces a distribution over trajectories $\xi = (o_1, a_1, \ldots, o_H, a_H)$ under any policy π, and a bounded return $R(\xi)$ or a binary success indicator. Accordingly, we write

$$S(x, \pi) \;:=\; \mathbb{E}\big[R(\xi) \,\big|\, \xi \sim \tau_x,\ \pi\big],$$

interpreting $S(x, \pi) \in [0, 1]$ as a success probability when R is an indicator, and as a normalized expected reward otherwise. We also use the complementary error notation $\mathrm{Err}(x, \pi) := 1 - S(x, \pi)$ (or an affine rescaling, depending on the evaluation convention).

A policy π maps the agent's information to actions. In the fully observed-context regime, we allow π to depend explicitly on x (i.e., $\pi(a_t \mid o_{1:t}, x)$). In

the latent-context regime, the policy depends only on observations (and internal memory), and any dependence on x is mediated by a learned inference mechanism. In both cases we posit the existence of an expert (or oracle) policy π_x^* that is optimal for τ_x under the same information structure available to the learner. The learned policy $\hat{\pi}$ is trained from offline episodes and is evaluated without further interaction during training; our primary concern is not within-task generalization, but rather the dependence of performance on the *set of contexts* represented in the dataset.

Training data. We assume access to an offline dataset $\mathcal{D}$ consisting of D episodes, each episode generated in some context $x_i \in \mathcal{X}$. The multiset of contexts appearing in the dataset is denoted by $\mathcal{S} := \{x_i\}_{i=1}^{D}$, and the set of distinct contexts by $U := \text{unique}(\mathcal{S})$, with cardinality $K := |U|$ (so $K \leq D$). We emphasize that D and K play different roles: D controls statistical estimation and optimization effects at a *fixed* context, whereas K (and more generally the geometry of U in $(\mathcal{X}, d)$) controls the ability to transfer to novel contexts. Since our goal is to isolate diversity effects, many of our statements later will be expressed as if an oracle learner could match the expert on each $x \in U$, thereby eliminating within-context estimation error; the remaining source of error is then attributable to the distance from a test context to its nearest representative in U.

Seen versus unseen evaluation. We fix a test distribution μ over contexts, intended to represent the deployment regime. For evaluation we consider a preregistered set of test contexts $\mathcal{X}_{\text{test}} \subseteq \mathcal{X}$ (often finite in empirical protocols), together with a partition into *seen* and *unseen* subsets:

$$\mathcal{X}_{\text{test}} = \mathcal{X}_{\text{seen}} \ \cup \ \mathcal{X}_{\text{unseen}}, \qquad \mathcal{X}_{\text{seen}} \cap \mathcal{X}_{\text{unseen}} = \emptyset,$$

where typically $\mathcal{X}_{\text{seen}} \subseteq U$ and $\mathcal{X}_{\text{unseen}} \cap U = \emptyset$. We report aggregate performance by integrating (or averaging) over the corresponding parts of the test distribution, e.g.,

$$\text{Err}_{\text{unseen}}(\hat{\pi}) \ := \ \mathbb{E}_{x\sim\mu}\big[\text{Err}(x, \hat{\pi}) \,\big|\, x \in \mathcal{X}_{\text{unseen}}\big],$$

with the analogous definition for $\text{Err}_{\text{seen}}(\hat{\pi})$. When $\mathcal{X}_{\text{test}}$ is finite and each context is evaluated by a bounded number of rollouts, we view the resulting binomial or Monte Carlo variability as separate from the population quantity above; it may be controlled by increasing evaluation rollouts and is orthogonal to the diversity question.

Coverage radius and effective diversity. The central geometric quantity we associate to a dataset is its *coverage radius* over the test set:

$$r(\mathcal{S}) \ := \ \sup_{x\in\mathcal{X}_{\text{test}}} \ \min_{x'\in U} d(x, x').$$

This is the largest distance from any test context to its nearest training context (among the distinct contexts present in the dataset). The dependence on U rather than on the full multiset $\mathcal{S}$ reflects the fact that duplicating a context does not improve coverage. In contexts where $\mathcal{X}_{\text{test}}$ is replaced by the support of μ, one obtains the corresponding distributional variant $r_\mu(\mathcal{S}) = \inf\{r : \Pr_{x\sim\mu}(\min_{x'\in U} d(x, x') \le r) = 1\}$; for clarity we focus on the test-set formulation used in typical benchmarks.

To convert radius into a scalar notion comparable across task spaces with different intrinsic dimension, we define the *effective diversity*

$$D_{\text{eff}}(\mathcal{S}) \;:=\; r(\mathcal{S})^{-d_{\mathcal{X}}},$$

where $d_{\mathcal{X}}$ denotes the doubling dimension (or another intrinsic metric dimension) of $(\mathcal{X}, d)$ under μ. This definition is motivated by the scaling of covering numbers in doubling spaces: if $N(\varepsilon)$ denotes the size of a minimal ε-net, then typically $N(\varepsilon) \asymp \varepsilon^{-d_{\mathcal{X}}}$. Thus $r(\mathcal{S})^{-d_{\mathcal{X}}}$ may be read as "the number of contexts needed to achieve radius $r(\mathcal{S})$," even when the observed K is not directly comparable across domains or when contexts are sampled non-uniformly.

Observed and latent contexts. In some domains the context x is explicit (e.g., a symbolic goal specification or a simulator seed), and d can be defined directly (e.g., edit distance over goal graphs, Hamming distance over discrete attributes, or Euclidean distance over continuous parameters). In other domains x is latent and must be inferred from observations. In this case we assume access to an embedding $\hat{\phi}(x)$ (learned or engineered) and define the operational metric by $d(x, x') := \|\hat{\phi}(x) - \hat{\phi}(x')\|$ or a related distance in embedding space. Any mismatch between d and the "true" task similarity relevant for transfer is treated as part of an irreducible system effect, and will appear later as an additive term in performance bounds. This separation allows us to state results in terms of a user-specified metric while making explicit that poor metric choice can limit transfer even when coverage is large.

This formalization reduces unseen-task generalization to a question about how well the training contexts cover the test contexts in $(\mathcal{X}, d)$. In the next section we show that, under a Lipschitz transfer hypothesis, $r(\mathcal{S})$ controls unseen error up to additive system terms, thereby motivating D_{eff} as the relevant scaling variable for unseen-task performance.

4 Main Theoretical Results: Upper Bounds

We now state an upper bound showing that, once the learner matches the expert on the contexts present in the dataset, the remaining degradation

on unseen contexts is controlled by the dataset coverage radius. The argument is intentionally modular: it separates (i) a geometric *transfer* term determined by $r(\mathcal{S})$, from (ii) a *within-context* learning term (approximation/optimization/statistical error) on the contexts actually observed, and (iii) an irreducible *system* term capturing metric mismatch, partial observability, and sim2real effects. In particular, the bound clarifies why increasing the raw episode count D can have negligible effect on $\mathrm{Err}_{\mathrm{unseen}}$ when it does not reduce $r(\mathcal{S})$.

Lipschitz transfer hypothesis. Fix a metric d on $\mathcal{X}$. We assume that success probabilities vary smoothly across contexts in the following sense: if a policy behaves as the expert for a nearby context, then its success at the target context cannot change too abruptly as a function of distance. Formally, we posit an $L < \infty$ such that for any $x, x' \in \mathcal{X}$ and any policy π satisfying $\pi(\cdot \mid \cdot, x') = \pi^*_{x'}(\cdot \mid \cdot, x')$ (i.e., it matches the expert when executed in context x'), one has

$$\left|S(x, \pi) - S(x', \pi^*_{x'})\right| \;\le\; L\, d(x, x'). \tag{1}$$

This hypothesis is weaker than requiring the dynamics of τ_x to be Lipschitz in x; it directly constrains the induced task-level performance functional. When x is latent and we use an embedding $\hat{\phi}(x)$ to define d, any violation of (1) is understood as contributing to the system term introduced below.

Seen-context learning guarantee. Let $U = \mathrm{unique}(\mathcal{S})$ denote the distinct contexts in the dataset. For each $x' \in U$ we measure how well the learned policy $\hat{\pi}$ matches the corresponding expert on that context. We allow a generic additive bound

$$S(x', \hat{\pi}) \;\ge\; S(x', \pi^*_{x'}) - \epsilon_{\mathrm{seen}}(x'), \qquad x' \in U, \tag{2}$$

where $\epsilon_{\mathrm{seen}}(x')$ aggregates approximation error of the policy class, optimization error of the training procedure, and finite-sample estimation error due to having only finitely many episodes at x'. In the oracle (realizable) regime emphasized for isolating diversity effects, we set $\epsilon_{\mathrm{seen}}(x') \le \epsilon_{\mathrm{sys}}$ uniformly; more generally we define $\epsilon_{\mathrm{learn}} := \sup_{x' \in U} \epsilon_{\mathrm{seen}}(x')$.

Theorem 4.1 (Unseen error upper bound via coverage radius)**.** *Assume* (1) *and* (2)*. Then for any test context $x \in \mathcal{X}_{\mathrm{test}}$,*

$$\mathrm{Err}(x, \hat{\pi}) \;\le\; \epsilon_{\mathrm{learn}} \;+\; \epsilon_{\mathrm{sys}} \;+\; L \min_{x' \in U} d(x, x'). \tag{3}$$

Consequently,

$$\mathrm{Err}_{\mathrm{unseen}}(\hat{\pi}) \;\le\; \epsilon_{\mathrm{learn}} \;+\; \epsilon_{\mathrm{sys}} \;+\; L\, r(\mathcal{S}). \tag{4}$$

Proof sketch. Fix $x \in \mathcal{X}_{\text{test}}$ and choose $x^\star \in \arg\min_{x' \in U} d(x, x')$. By (2) we control the performance of $\hat{\pi}$ at $x^\star$ relative to $\pi^*_{x^\star}$. To relate x to $x^\star$, consider the policy $\pi^*_{x^\star}$, which by definition matches itself on context $x^\star$; applying (1) yields

$$S(x, \pi^*_{x^\star}) \;\geq\; S(x^\star, \pi^*_{x^\star}) - L\, d(x, x^\star).$$

Finally, we use that $\hat{\pi}$ is within $\epsilon_{\text{learn}} + \epsilon_{\text{sys}}$ of $\pi^*_{x^\star}$ at $x^\star$, and we pessimistically transfer this deviation to x by absorbing any remaining mismatch into ϵ_{sys}. Rearranging from success to error yields (3). Taking an average (or supremum) over $x \in \mathcal{X}_{\text{unseen}}$ gives (4), and using the definition of $r(\mathcal{S})$ completes the argument. □

Interpretation and decomposition. The bound (4) exhibits the promised three-way decomposition:

$$\text{Err}_{\text{unseen}} \;\lesssim\; \underbrace{L\, r(\mathcal{S})}_{\text{coverage/transfer}} \;+\; \underbrace{\epsilon_{\text{learn}}}_{\text{approx. \& opt. on seen}} \;+\; \underbrace{\epsilon_{\text{sys}}}_{\text{irreducible system}}.$$

The coverage term depends only on the *set* of represented contexts U and is unchanged by repeating contexts; thus increasing D while keeping $K = |U|$ fixed can reduce ϵ_{learn} (via better estimation on seen contexts) but cannot improve the leading transfer term. Conversely, increasing K in a way that reduces $r(\mathcal{S})$ improves the bound even if the number of episodes per context is held fixed.

Effective diversity as the scaling variable. In a doubling space, $r(\mathcal{S})$ is the correct geometric bottleneck, and $D_{\text{eff}}(\mathcal{S}) = r(\mathcal{S})^{-d_{\mathcal{X}}}$ provides a dimension-normalized proxy for how many contexts have effectively been covered. Writing (4) in terms of D_{eff} yields

$$\text{Err}_{\text{unseen}}(\hat{\pi}) \;\leq\; \epsilon_{\text{learn}} + \epsilon_{\text{sys}} + L\, D_{\text{eff}}(\mathcal{S})^{-1/d_{\mathcal{X}}},$$

which motivates fitting a power law in D_{eff} rather than in D. The subsequent lower bound results will show that, without additional structure beyond Lipschitz transfer, dependence on $r(\mathcal{S})$ (and hence on D_{eff}) is unavoidable up to constant and logarithmic factors.

5 Main Theoretical Results: Lower Bounds

We now complement the coverage-based upper bound by a minimax lower bound showing that, under no assumptions beyond the Lipschitz transfer hypothesis, dependence on the coverage radius is information-theoretically unavoidable. The guiding point is that the dataset reveals behavior only on the represented contexts. If there exists a test context at distance $\approx r(\mathcal{S})$ from all represented contexts, then an adversary can modify the task family

in a way that is (i) consistent with all training episodes, (ii) still L-Lipschitz across contexts, yet (iii) forces any learned policy to incur error at least proportional to that distance on at least one unseen context. Consequently, any improvement in unseen performance must come from reducing $r(\mathcal{S})$ (equivalently increasing D_{eff}), rather than from algorithmic ingenuity alone.

Theorem 5.1 (Minimax lower bound: coverage is necessary)**.** *Fix a (multi)set of training contexts $\mathcal{S} \subset \mathcal{X}$ and let $U = \text{unique}(\mathcal{S})$. There exists a context-indexed task family $\{\tau_x\}_{x\in\mathcal{X}}$ with bounded success probabilities, satisfying the Lipschitz transfer hypothesis* (1)*, such that for any (possibly randomized) learning algorithm $\mathcal{A}$ that outputs a policy $\hat{\pi} = \mathcal{A}(\mathcal{D})$ using only episodes collected on contexts in $\mathcal{S}$, we have*

$$\sup_{x\in\mathcal{X}_{\text{test}}} \text{Err}(x, \hat{\pi}) \;\geq\; c\, r(\mathcal{S}) \;-\; \epsilon_{\text{obs}}, \tag{5}$$

for a universal constant $c > 0$ and an additive term ϵ_{obs} capturing any exogenous observation noise or non-identifiability that persists even at a fixed context.

Proof idea (indistinguishability via a packing argument). We outline a standard reduction (Le Cam / Yao) adapted to our context-indexed setting. Let $r := r(\mathcal{S})$. By definition of r, there exists $x_0 \in \mathcal{X}_{\text{test}}$ such that $d(x_0, u) \geq r$ for all $u \in U$. We will construct two task families, $\mathcal{T}^+$ and $\mathcal{T}^-$, which agree on *all* contexts in U (hence generate identical training data distributions on $\mathcal{S}$), but induce different optimal behavior at x_0 in a way that is detectable only by interacting with contexts near x_0. Since the learner never observes episodes at x_0 (or near it, if r is large), it cannot reliably distinguish $\mathcal{T}^+$ from $\mathcal{T}^-$, implying that its output must be suboptimal on at least one of them at x_0.

Concretely, consider a stylized family of horizon-H tasks in which the only decision that matters is a single binary action $a \in \{+1, -1\}$ taken at the first step; the episode then terminates with a Bernoulli success whose mean depends on (x, a). (This can be realized as a degenerate POMDP with trivial observations and a terminal reward.) For a chosen center x_0, define a *bump* function

$$b(x) \;:=\; \max\Big\{0,\; 1 - \frac{d(x, x_0)}{r}\Big\},$$

which is $1/r$-Lipschitz with respect to d and satisfies $b(u) = 0$ for all $u \in U$ since $d(u, x_0) \geq r$. Now define two instances $\mathcal{T}^+$ and $\mathcal{T}^-$ by specifying success probabilities

$$S_{\pm}(x, a) \;:=\; \frac{1}{2} \;\pm\; \frac{Lr}{4}\, b(x) \cdot a,$$

clipped to $[0, 1]$ if desired (for small enough r this clipping is unnecessary; otherwise one may rescale constants). Several properties are immediate.

First, for every training context $u \in U$ we have $b(u) = 0$, hence $S_+(u, a) = S_-(u, a) = 1/2$ for both actions. Therefore, the entire distribution over training episodes (including actions, observations, and outcomes) is identical under $\mathcal{T}^+$ and $\mathcal{T}^-$, regardless of the data-collection policy used on $\mathcal{S}$. Second, at the unseen context x_0 we have $b(x_0) = 1$, and the optimal action differs between the two instances: $\mathcal{T}^+$ prefers $a = +1$ while $\mathcal{T}^-$ prefers $a = -1$, with an optimality gap of order Lr. Third, the mapping $x \mapsto S_\pm(x, a)$ is L-Lipschitz up to constant factors, since b is $1/r$-Lipschitz and the prefactor is proportional to Lr. This realizes the transfer regularity demanded by (1) at the level of the induced success functional.

Under this construction, any learning algorithm $\mathcal{A}$ produces (possibly randomized) $\hat{\pi}$ based on data that is identically distributed under $\mathcal{T}^+$ and $\mathcal{T}^-$. Thus, $\hat{\pi}$ cannot correlate with the hidden sign. By a two-point testing argument, the expected suboptimality of $\hat{\pi}$ at x_0 under the uniform prior over $\{\mathcal{T}^+, \mathcal{T}^-\}$ is at least a constant fraction of the gap, hence at least $\Omega(Lr)$. Converting suboptimality in success to an error lower bound yields (5) (absorbing constants into c and any unavoidable ambiguity into ϵ_{obs}). A more refined version replaces a single x_0 by a packing of many well-separated centers and uses Fano's inequality; this yields the same linear dependence on $r(\mathcal{S})$ while making explicit that the adversary can place the "hard" region anywhere that remains uncovered by U.

Tightness and logarithmic factors. Theorem 5.1 matches the upper bound dependence on $r(\mathcal{S})$ up to constants (and the additive system/observation terms). In particular, in regimes where ϵ_{sys} and ϵ_{obs} are negligible, the quantity $r(\mathcal{S})$ is the correct rate-determining bottleneck: no algorithm can guarantee $\text{Err}_{\text{unseen}} = o(r(\mathcal{S}))$ uniformly over all task families obeying Lipschitz transfer. When $\mathcal{S}$ is itself random (e.g., K i.i.d. distinct contexts), the remaining gap between achievable and unavoidable rates arises not from learning but from geometry and concentration: expected coverage radii in doubling spaces typically scale as $K^{-1/d_{\mathcal{X}}}$ up to $\log K$ factors stemming from uniform control over an ϵ-net. Thus, while the next section will translate coverage into a power law in D_{eff}, the present lower bound already establishes that any such power law must ultimately be governed by the geometry of context coverage rather than the raw episode count.

6 From Coverage to Power Laws: Doubling-Dimension Radius Scaling

We now translate the radius-based upper and lower bounds into an explicit scaling prediction as the number of *distinct* training contexts grows. The key point is that, under the Lipschitz transfer model, the only dataset statistic

that enters the worst-case unseen error is the coverage radius

$$r(\mathcal{S}) \;:=\; \sup_{x \in \mathcal{X}_{\text{test}}} \min_{x' \in \mathcal{S}} d(x, x'),$$

where we may without loss restrict $\mathcal{S}$ to its set of unique contexts. Thus, to understand how unseen error improves with more data, it suffices to understand how quickly $r(\mathcal{S})$ decreases as we add new, previously-unseen contexts.

Radius scaling in doubling spaces. Assume that contexts are drawn i.i.d. from a test distribution μ supported on a metric space $(\mathcal{X}, d)$ with finite doubling dimension $d_{\mathcal{X}}$. Let $\mathcal{S}_K$ denote K i.i.d. draws from μ, and let $U_K = \text{unique}(\mathcal{S}_K)$ be the induced set of distinct contexts (we ignore repeats, since repeats do not improve coverage). Classical covering-number arguments imply that for sufficiently large K,

$$\mathbb{E}\big[r(U_K)\big] \;=\; \Theta\Big(\big(K/\log K\big)^{-1/d_{\mathcal{X}}}\Big), \qquad r(U_K) \;=\; \tilde{\Theta}\big(K^{-1/d_{\mathcal{X}}}\big) \;\text{ w.h.p.} \quad (6)$$

The intuition is standard. In a doubling space, the number of balls of radius ε required to cover typical mass scales as $\varepsilon^{-d_{\mathcal{X}}}$ up to constants. If we choose ε so that the covering number is on the order of K, then i.i.d. sampling populates most covering cells with high probability; the logarithmic factor arises from the uniform control needed to avoid leaving any cell empty (a coupon-collector effect over an ε-net). Equation (6) formalizes that geometric statement in the quantity we care about, namely the maximal distance from a test point to its nearest sampled context.

Unseen error as a power law in effective diversity. Combining (6) with the coverage-based bound (Theorem 1) yields an explicit prediction for the *unseen* scaling curve. Indeed, for a learner that is essentially optimal on the represented contexts up to ϵ_{sys}, we have

$$\text{Err}_{\text{unseen}} \;\le\; \epsilon_{\text{sys}} + L\, r(U_K).$$

Taking expectations and substituting (6) gives

$$\mathbb{E}\big[\text{Err}_{\text{unseen}}\big] \;\le\; \epsilon_{\text{sys}} \;+\; \tilde{O}\big(L\, K^{-1/d_{\mathcal{X}}}\big), \quad (7)$$

and Theorem 4 implies that this dependence on $r(U_K)$ (hence on $K^{-1/d_{\mathcal{X}}}$) is unavoidable up to constants in the worst case. It is therefore natural to reparametrize the horizontal axis by an *effective diversity*

$$D_{\text{eff}}(\mathcal{S}) \;:=\; r(\mathcal{S})^{-d_{\mathcal{X}}},$$

which is monotone in coverage and (in doubling spaces) is equivalent to the number of "ε-balls worth of support" covered by the dataset. Writing (7) in terms of D_{eff} yields the power-law form

$$\text{Err}_{\text{unseen}} \;\approx\; A\, D_{\text{eff}}^{\alpha_{\text{unseen}}} \;+\; E, \qquad \alpha_{\text{unseen}} \;=\; -\frac{1}{d_{\mathcal{X}}}, \qquad E \;\approx\; \epsilon_{\text{sys}}, \quad (8)$$

where A absorbs L and distribution-dependent constants and $\tilde{O}(\cdot)$ logarithms have been suppressed. The exponent α_{unseen} is stable in the sense that it depends only on the intrinsic metric dimension of the context space under μ, not on the learning architecture, optimizer, or other implementation details (which, in our abstraction, affect primarily the additive floor E).

Why unseen scaling is typically slower than seen scaling. The same dataset may exhibit substantially different scaling behavior on *seen* and *unseen* contexts. On seen contexts, additional episodes at already-represented contexts can reduce estimation error, imitation mismatch, and other context-local effects; consequently, seen performance may improve primarily with total episode count D, and its effective exponent can be steeper when the learner benefits from repeated supervision. By contrast, under the present transfer model, unseen performance cannot improve unless the set of represented contexts becomes a finer net over $\mathcal{X}_{\text{test}}$. Repeating an already-covered context does not reduce $r(\mathcal{S})$, hence cannot improve the worst-case unseen guarantee beyond the system floor. In this sense, the unseen exponent is "geometry-limited": it is controlled by $d_{\mathcal{X}}$ through (8), and in typical regimes $1/d_{\mathcal{X}}$ is small enough that $|\alpha_{\text{unseen}}| < |\alpha_{\text{seen}}|$.

Implications for dataset construction under a fixed episode budget. Let D be the total episode budget. If we allocate m episodes per context and collect K distinct contexts, then $D = Km$. Under the idealized assumption that the learner matches the expert on each represented context once that context is present (i.e., no residual estimation error from finite m), the bound depends only on K through $r(U_K)$, so the optimal strategy is to maximize K (take $m = 1$) and thereby minimize $r(\mathcal{S})$. In more realistic settings, finite m reduces within-context error but does not change coverage; thus we obtain an explicit design tension: increasing m can lower the additive terms bundled into E, while increasing K improves the geometric term $AD_{\text{eff}}^{-1/d_{\mathcal{X}}}$. Our framework separates these contributions and predicts when additional repeats will saturate (once E dominates) versus when acquiring new contexts will continue to pay off.

Finally, since D_{eff} is defined through a metric, the same analysis applies when contexts are not directly observed, provided we can work in a learned embedding $\hat{\phi}(x)$ and an induced distance $d(\hat{\phi}(x), \hat{\phi}(x'))$ that preserves neighborhood relations relevant for transfer. This motivates estimating D_{eff} directly from the observed set of contexts (or embeddings) and designing sampling procedures that explicitly minimize the empirical coverage radius. The next section makes these statements algorithmic by giving concrete estimators and near-optimal context-selection rules.

7 Algorithms: Estimating Effective Diversity and Designing High-Coverage Datasets

The preceding analysis reduces unseen-task behavior to a geometric statistic of the set of *distinct* contexts represented in the dataset. We now make this reduction operational: given a finite dataset of episodes (possibly with repeated contexts), we estimate the coverage radius and hence D_{eff}; given control over a procedural generator or a large candidate pool of contexts, we construct a high-coverage dataset subject to a fixed episode budget.

Estimating coverage radius and D_{eff} from a dataset. Let $\mathcal{D} = \{(x_i, \text{episode}_i)\}_{i=1}^{D}$ be an offline dataset and let

$$U \;:=\; \text{unique}(\{x_i\}_{i=1}^{D}), \qquad K \;:=\; |U|.$$

Since repeats do not decrease the nearest-neighbor distance to a test context, all coverage-based quantities depend on $\mathcal{D}$ only through U. In principle the radius is

$$r(U) \;=\; \sup_{x \in \mathcal{X}_{\text{test}}} \min_{u \in U} d(x, u),$$

but the supremum is inaccessible unless $\mathcal{X}_{\text{test}}$ is finite. We therefore work with an *evaluation pool* $\mathcal{X}_{\text{eval}} = \{x^{(j)}\}_{j=1}^{M}$ sampled from the fixed (preregistered) test distribution, and define the plug-in estimator

$$\hat{r} \;:=\; \max_{1 \le j \le M} \min_{u \in U} d\big(x^{(j)}, u\big), \qquad \widehat{D}_{\text{eff}} \;:=\; (\hat{r} + \delta)^{-\hat{d}}, \tag{9}$$

where $\delta > 0$ is a numerical stabilizer and $\hat{d}$ is an estimate of $d_{\mathcal{X}}$. The estimator (9) is conservative in the sense that it upper bounds the empirical radius on $\mathcal{X}_{\text{eval}}$ exactly; when $\mathcal{X}_{\text{eval}}$ is an i.i.d. sample from μ, standard uniform convergence arguments imply that $\hat{r}$ concentrates around the population radius at a rate governed by M and the metric entropy of $(\mathcal{X}, d)$, which in our setting is controlled by $d_{\mathcal{X}}$.

Metrics from discrete factors and hybrid context descriptions. In many benchmarks the context admits an explicit factorization, e.g. $x = (\text{object types}, \text{layout}, \text{goal})$. A direct choice is a weighted Hamming or edit-type metric

$$d(x, x') \;:=\; \sum_{\ell=1}^{p} w_\ell \, \mathbf{1}\{x_\ell \neq x'_\ell\},$$

where weights w_ℓ encode which factors are believed to drive transfer. Such metrics make D_{eff} easy to compute and interpret: $\min_{u \in U} d(x, u)$ is simply the number (or weighted number) of factor mismatches to the nearest seen configuration. When factors are partially ordered (e.g. counts of distractors),

one can replace the indicator by an absolute difference. The framework is agnostic to these choices; the role of the metric is only to formalize which context perturbations should be "small" for policy transfer.

Embedding-based estimation when context is latent or high-dimensional. If contexts are not directly observed, we assume we can compute an embedding $\hat{\phi}(x) \in \mathbb{R}^m$ (from metadata, images, or a learned encoder) and use an induced distance

$$d_\phi(x, x') \;:=\; \big\|\hat{\phi}(x) - \hat{\phi}(x')\big\|_2 \quad \text{or} \quad d_\phi(x, x') \;:=\; 1 - \frac{\langle \hat{\phi}(x), \hat{\phi}(x') \rangle}{\|\hat{\phi}(x)\|\,\|\hat{\phi}(x')\|}.$$

We then replace d by d_ϕ in (9). This substitution is justified whenever $\hat{\phi}$ is *neighborhood-preserving* for transfer-relevant variations, i.e. Lipschitz transfer holds with respect to d_ϕ (possibly with a different constant). Practically, we recommend sanity checks that the induced nearest-neighbor structure aligns with empirical transfer: if a held-out context x is close (in d_ϕ) to some $u \in U$, then the policy trained on u should succeed on x at an elevated rate relative to distant u'.

Constructing high-coverage datasets: farthest-first (metric k-center). When we can choose which contexts to collect episodes from, the natural objective is to minimize the coverage radius subject to a budget of K distinct contexts:

$$\min_{U \subseteq \mathcal{C},\, |U| = K} \; \max_{x \in \mathcal{X}_{\text{eval}}} \; \min_{u \in U} d(x, u),$$

where $\mathcal{C}$ is a large candidate set produced by the generator (or an on-the-fly sampler). This is precisely the metric k-center problem, which is NP-hard in general, so we adopt the classical farthest-first traversal: start from an arbitrary seed $u_1 \in \mathcal{C}$ and iteratively add

$$u_t \;\in\; \arg\max_{c \in \mathcal{C}} \; \min_{u \in U_{t-1}} d(c, u), \qquad U_t := U_{t-1} \cup \{u_t\}.$$

The algorithm greedily decreases the maximum uncovered distance on the candidate pool and admits a worst-case 2-approximation guarantee: the radius achieved by U_K is at most twice the optimum achievable with K centers on the same pool. Composed with the coverage-to-error relationship developed earlier, this yields a near-optimal design rule for minimizing the worst-case unseen bound up to the universal factor 2 (and additive ϵ_{sys}).

Episode allocation: separating coverage from repeats. Given a total episode budget D, we may choose K contexts and allocate $m = D/K$ episodes per context. Coverage-based terms depend primarily on K (through

$r(U)$), whereas within-context estimation and behavior cloning noise decrease with m. This suggests a practical two-stage design: (i) select K contexts using farthest-first to minimize $\hat{r}$; (ii) allocate remaining episodes adaptively to contexts where imitation loss is highest, without changing U. This explicitly targets the decomposition "geometric error plus system/local error" that motivates our scaling law.

Computational complexity and implementation notes. Computing $\hat{r}$ on an evaluation pool of size M against K selected contexts costs $O(MK)$ distance evaluations, which is typically small compared to policy training. Farthest-first selection over a candidate pool $\mathcal{C}$ of size N can be implemented in $O(NK)$ time by maintaining, for each candidate c, its current nearest-center distance $\min_{u \in U_t} d(c, u)$ and updating this value incrementally. Both procedures are compatible with approximate nearest-neighbor search and with streaming candidates (where $\mathcal{C}$ is too large to store) by maintaining a reservoir of promising farthest points. These algorithmic choices are the minimal machinery needed to turn the geometric viewpoint into concrete experimental protocols, which we specify next.

8 Experimental Design (Recommended)

Our empirical goal is to isolate the geometric quantity D_{eff} as the driver of generalization to $\mathcal{X}_{\text{unseen}}$, and to separate it from effects attributable to raw episode count, representation choice, and evaluation noise. To this end we recommend experiments that, using a procedural generator or a large precomputed pool, can vary *count without diversity* and *diversity without count* while holding the evaluation protocol fixed.

Two independent knobs: repeats versus new contexts. Fix a total episode budget D and a target number of distinct contexts K. We construct datasets of the form $\mathcal{D}(K, m)$ with $m := \lfloor D/K \rfloor$ episodes per context, where the context set U is chosen either (i) i.i.d. from the generator distribution, or (ii) by a coverage-maximizing rule (e.g. farthest-first on a candidate pool). To vary *count without diversity*, we fix K and increase m (hence D) by collecting additional episodes on the same U. To vary *diversity without count*, we fix D and increase K (hence decrease m), regenerating U each time. Under the coverage model, the unseen term should primarily track K (equivalently $\widehat{D}_{\text{eff}}$), whereas additional repeats should manifest mainly through reductions in within-context behavioral cloning error and through a decreased effective ϵ_{sys} (e.g. better state coverage, reduced stochasticity), without improving the radius term.

Procedural generators with controllable intrinsic dimension. When possible, we recommend designing generator families with explicit factor structure and adjustable "intrinsic dimension." Concretely, let $x = (x_1, \ldots, x_p)$ with each factor corresponding to a semantic degree of freedom (object identity, distractor count, room topology, goal specification). By choosing which factors are allowed to vary, and by adjusting their cardinalities or ranges, we can create regimes with different effective $d_{\mathcal{X}}$ while keeping the perceptual channel and action space unchanged. This enables a direct test of the prediction that the unseen exponent $\alpha_{\text{unseen}} \approx -1/d_{\mathcal{X}}$ becomes less negative as the context family becomes higher-dimensional.

Preregistered evaluation pool and estimator logging. We assume a fixed unseen evaluation pool $\mathcal{X}_{\text{eval}} \subset \mathcal{X}_{\text{unseen}}$ sampled once from μ and reused across all conditions. For each trained policy we evaluate n_{eval} rollouts per context and report the empirical unseen error

$$\widehat{\text{Err}}_{\text{unseen}} \;:=\; 1 - \frac{1}{|\mathcal{X}_{\text{eval}}|} \sum_{x \in \mathcal{X}_{\text{eval}}} \widehat{S}(x, \hat{\pi}),$$

together with binomial confidence intervals aggregated over contexts. Simultaneously, we compute and log $\hat{r}$ and $\widehat{D}_{\text{eff}}$ from the training set (using the same $\mathcal{X}_{\text{eval}}$ for the plug-in radius estimate), as well as K and m. This makes it possible to regress performance against $(D, K, \widehat{D}_{\text{eff}})$ rather than only against D, and to diagnose failures of curve collapse.

Predicted curve collapse and scaling fits. For each dataset condition we fit a parametric form

$$\widehat{\text{Err}}_{\text{unseen}} \;\approx\; \Big(\frac{A}{\widehat{D}_{\text{eff}}}\Big)^{\alpha_{\text{unseen}}} + E,$$

with $E \geq 0$ capturing irreducible error, and compare it to the analogous fit against raw episode count D. The central prediction is that plotting $\widehat{\text{Err}}_{\text{unseen}}$ versus $\widehat{D}_{\text{eff}}$ yields substantially reduced variance across generator settings and across dataset construction methods (random versus farthest-first), whereas plotting against D does not. A complementary diagnostic is a two-way ablation: (i) fix K and vary D via repeats, and (ii) fix D and vary K. The model predicts that regime (ii) produces a clear monotone improvement on unseen contexts, while regime (i) produces at best a weak improvement that saturates quickly once within-context imitation error is negligible.

Architecture ablations: VLA vs. VLM vs. PVR. To test that D_{eff} captures a geometric limitation rather than an architecture-specific artifact, we recommend repeating the above scaling study for distinct policy families:

(a) a vision-language-action (VLA) policy trained end-to-end by behavior cloning; (b) a vision-language model (VLM) used for high-level inference combined with a fixed low-level controller or planner; and (c) a policy with a pretrained visual representation (PVR) and a smaller action head. Our hypothesis is that these choices primarily shift (A, E) (through representation quality, optimization, and partial observability) while leaving the fitted exponent α_{unseen} approximately invariant within a fixed context family and metric. Deviations are informative: if one architecture changes α_{unseen} materially, this suggests that the relevant metric for transfer differs (e.g. a representation induces a different neighborhood structure), or that the Lipschitz assumption is violated in a way that interacts with the model class.

Metric and embedding sanity checks. Finally, we recommend an explicit metric-ablation protocol: compute $\widehat{D}_{\text{eff}}$ under multiple plausible distances (factor-weighted Hamming, learned embedding distance, hybrid metrics) and compare which distance yields the tightest collapse of unseen scaling. A minimal sanity check is nearest-neighbor transfer: for held-out contexts x, success should correlate with $\min_{u \in U} d(x, u)$ computed under the chosen metric. If this correlation is absent, then the empirical failure should be attributed to a mis-specified geometry rather than to the coverage principle itself.

9 Discussion and Limitations

Our central claim is conditional: if transfer is controlled by a metric geometry on contexts, and if the learner can essentially match the expert on the training contexts up to an additive term, then unseen performance is governed by coverage radius (equivalently D_{eff}). The same conditionality delineates the main failure modes. We therefore record where the Lipschitz–coverage account can break, which quantities become ill-defined, and what diagnostic signatures to expect.

When Lipschitz transfer fails (non-smooth task families and discontinuities). Assumption (A1) posits that success varies at most linearly with context distance under a coupling that compares π to an expert $\pi^*_{x'}$ on x'. Many families violate this in structurally unavoidable ways. First, tasks may exhibit *phase transitions*: a small context change induces a qualitative change in required strategy (e.g. a key switches location from reachable to unreachable, a door becomes locked, or a single distractor triggers a different instruction parse). In such cases $S(x, \pi^*_{x'})$ can drop abruptly even when $d(x, x')$ is small for natural choices of d, implying an effectively unbounded local Lipschitz constant. Second, the relevant geometry may be *non-metric* for the agent: two contexts can be perceptually aliased under the observation

channel, so that $d(x, x')$ is small in latent semantics but indistinguishable in observations (or conversely, visually similar but semantically far). In either direction, the nearest-neighbor surrogate $\min_{x' \in \mathcal{S}} d(x, x')$ ceases to predict success. Empirically, this manifests as weak or absent correlation between held-out success and nearest-training distance under any candidate metric; in that regime, improving $r(\mathcal{S})$ will not reliably improve unseen performance, and one should attribute the failure to mis-specified geometry rather than to insufficient diversity per se.

Representation learning and the choice of metric. Our theorems are stated in terms of a metric d on contexts. In realistic settings, contexts are not directly observed, and we instead compute distances using an embedding $\hat{\phi}(x)$ learned from raw perceptual streams or metadata. This introduces two distinct errors. (i) *Metric distortion:* if $\|\hat{\phi}(x) - \hat{\phi}(x')\|$ does not preserve the neighborhood structure relevant for transfer, then $\widehat{D}_{\text{eff}}$ can be spuriously large (contexts appear spread out) or spuriously small (contexts collapse), and the fitted exponent can drift. (ii) *Embedding estimation noise:* if x is inferred from trajectories, then $\hat{\phi}(x)$ depends on policy-induced observations, producing a feedback loop: better policies yield better context estimates, which in turn change the measured coverage. A conservative remedy is to compute $\hat{r}$ on a preregistered evaluation pool using a context descriptor that is independent of the learned policy (e.g. generator parameters), and to treat embedding-based $\widehat{D}_{\text{eff}}$ as a secondary analysis. More generally, if the representation is part of the learned system, then the effective Lipschitz constant L and even the intrinsic dimension $d_{\mathcal{X}}$ become *representation-dependent*, and it is not meaningful to compare α_{unseen} across models without specifying the induced geometry.

Covariate shift beyond context and compounding in POMDPs. The coverage bound controls transfer across *tasks indexed by* x, but it does not by itself control distribution shift in *state visitation* within a fixed x. In offline imitation, a policy trained on expert data may visit states absent from the dataset even on seen contexts, creating error cascades that are not captured by $r(\mathcal{S})$. One may attempt to absorb these effects into ϵ_{sys}, but doing so obscures an important distinction: increasing repeats m at fixed K can reduce such covariate-shift effects by improving within-context coverage of state space, whereas increasing K at fixed D can worsen them by making each context under-sampled. Consequently, the prediction "unseen improves mainly with K" should be interpreted as holding in a regime where within-context imitation error is already small. A practical diagnostic is to measure seen-context performance as m increases: if seen error does not saturate, then the regime is not yet coverage-limited, and one should not expect a clean D_{eff} collapse on unseen contexts.

Intrinsic dimension and finite-sample effects. The exponent relation $\alpha_{\text{unseen}} \approx -1/d_{\mathcal{X}}$ is asymptotic and distributional. In finite samples, estimates of $d_{\mathcal{X}}$ (e.g. via doubling tests or slope of covering numbers) can be unstable, and $\widehat{D}_{\text{eff}} = (\hat{r} + \delta)^{-\hat{d}}$ compounds error multiplicatively. Moreover, Theorem 2 includes logarithmic factors and implicitly assumes that μ is not concentrated on a lower-dimensional manifold except through $d_{\mathcal{X}}$; heavy-tailed or multimodal context distributions can yield different effective rates. In such cases, we recommend reporting $\hat{r}$ directly alongside $\widehat{D}_{\text{eff}}$, since $r(\mathcal{S})$ is the quantity that appears in the upper and lower bounds without requiring dimension estimation.

Connections to compute-optimal scaling and inference-time compute. Our analysis holds computation fixed and places all irreducible effects into ϵ_{sys}. This is a limitation when comparing systems with different training or inference compute. Increased training compute can reduce optimization error and representation error, effectively decreasing ϵ_{sys} and possibly reducing the empirical L by learning features that linearize transfer. Inference-time compute (planning, search, tool use, or test-time adaptation) changes the picture more sharply: it can induce a policy class that is *not* well-modeled as a single static π, and can improve success on contexts far from $\mathcal{S}$ without changing $r(\mathcal{S})$ by leveraging additional structure at test time. From our perspective, such mechanisms either (i) lower ϵ_{sys} by correcting partial observability and exploration failures, or (ii) replace the effective metric by one in which task difficulty varies more smoothly. Thus, compute-optimal frontiers should be stated in terms of triples $(D_{\text{eff}}, \text{train compute}, \text{inference compute})$, with D_{eff} accounting for geometric coverage and compute accounting for the residual term.

Summary of what remains outside the coverage model. We view D_{eff} as necessary and often predictive, but not sufficient: it cannot certify performance when the task family is discontinuous, when the induced geometry is misspecified, or when POMDP covariate shift dominates. These limitations suggest an empirical workflow: first verify nearest-neighbor distance predictivity; next confirm that seen-context error is near its floor; only then interpret unseen scaling primarily through $\widehat{D}_{\text{eff}}$ rather than through raw episode count.

10 Conclusion

We have isolated a single geometric quantity—the coverage radius

$$r(\mathcal{S}) := \sup_{x \in \mathcal{X}_{\text{test}}} \min_{x' \in \mathcal{S}} d(x, x')$$

—that mediates, up to the Lipschitz constant and an additive system term, the achievable error on unseen contexts when training data are collected on a set (or multiset) of contexts $\mathcal{S}$. The upper and lower bounds together imply that, within the stated hypotheses, one cannot generally trade away context coverage by collecting more repetitions on already-seen contexts, nor can one generally beat the dependence on $r(\mathcal{S})$ without strengthening structural assumptions beyond Lipschitz transfer. This yields a simple organizing principle: for generalization across tasks indexed by x, *the relevant sample size is geometric* and is better summarized by an effective diversity

$$D_{\text{eff}}(\mathcal{S}) \;:=\; r(\mathcal{S})^{-d_{\mathcal{X}}}$$

than by the raw episode count D.

This perspective provides a unifying explanation for a recurring empirical observation: unseen-task scaling exponents are often weak (i.e. $|\alpha_{\text{unseen}}|$ is small) relative to seen-task exponents. Under doubling-dimension assumptions, $r(\mathcal{S})$ decreases like a negative power of the number of *distinct* contexts, rather than the number of episodes, and the corresponding exponent satisfies $\alpha_{\text{unseen}} \approx -1/d_{\mathcal{X}}$ (up to logarithmic factors and ϵ_{sys} floors). In particular, when $d_{\mathcal{X}}$ is moderate or large, the geometric rate is intrinsically slow, so doubling D while keeping K (the number of distinct contexts) fixed is predicted to yield limited improvement on unseen contexts once within-context imitation error has saturated. Conversely, two datasets with similar D but very different $r(\mathcal{S})$ should exhibit markedly different unseen performance. We therefore expect that replacing D by $\widehat{D}_{\text{eff}}$ (or reporting $\hat{r}$ directly) will reduce variance across laboratories, model classes, and data-collection protocols whenever the transfer geometry is approximately stable.

The practical implication for 2026 "data engines" is that context selection should be treated as a first-class optimization variable rather than an incidental byproduct of logging. Concretely, a data engine can (i) maintain a candidate pool $\mathcal{C}$ of contexts (e.g. simulator seeds, procedural-generator parameters, or task descriptors), (ii) define a metric d on these descriptors (or a learned embedding used only for selection), and (iii) allocate new distinct contexts by a k-center approximation such as farthest-first traversal. This directly targets the quantity that appears in the transfer bound. When the episode budget is $D = K \cdot m$, the engine should separate two regimes: increase m until performance on already-collected contexts is near its floor, and then prioritize increasing K to decrease $r(\mathcal{S})$. In this view, "dataset size" is a two-dimensional budget (distinct contexts versus repetitions), and the correct control knob for unseen generalization is coverage.

For benchmarks and shared evaluations, we advocate a corresponding shift in what is measured and disclosed. A benchmark that claims to test generalization across contexts should specify (at minimum) (a) the context space $\mathcal{X}$ or a canonical descriptor space, (b) a preregistered unseen evaluation

set $\mathcal{X}_{\text{unseen}}$, and (c) a metric d used for coverage reporting. Participants should be required to report K, $\hat{r}$ computed on the fixed evaluation pool, and the context-selection procedure that produced $\mathcal{S}$. Such reporting makes it possible to distinguish improvements attributable to broader task coverage from improvements attributable to algorithmic advances (e.g. reducing ϵ_{sys}). It also enables apples-to-apples comparisons when different teams allocate budgets differently: a method that attains lower unseen error at the same $\hat{r}$ is plausibly improving representation, optimization, inference-time reasoning, or robustness, whereas a method that attains lower unseen error primarily by lowering $\hat{r}$ is primarily improving data coverage.

Finally, the coverage formalism suggests a principled way to design scaling studies that remain interpretable as models and environments evolve. Rather than plotting unseen error versus D alone, one should (i) fit unseen scaling laws in terms of $\widehat{D}_{\text{eff}}$ or $\hat{r}$, (ii) explicitly model floors via ϵ_{sys}, and (iii) stratify by intrinsic dimension estimates when feasible. This yields a testable prediction: if two systems share comparable geometry (similar induced $d_{\mathcal{X}}$ and L) and comparable irreducible error, then unseen performance should collapse as a function of $\widehat{D}_{\text{eff}}$ even when their raw episode counts differ. When the collapse fails, the correct conclusion is not merely that "scaling is noisy," but that at least one of the modeling ingredients—geometry, realizability on seen contexts, or the interpretation of ϵ_{sys}—has changed.

In summary, we propose effective diversity as a common language linking dataset design, benchmark construction, and scaling-law interpretation for generalization across task contexts. Within the Lipschitz–coverage model, weak unseen exponents are not anomalous; they are the expected consequence of finite intrinsic dimension and incomplete coverage. The actionable response is correspondingly geometric: build data engines that actively minimize coverage radius, and build benchmarks that measure and report it.