Bounding Global and Local Compression Error of Signal Parameterizations

Quang Luong Nhat Nguyen **Sara Fridovich-Keil**
School of Electrical and Computer Engineering
Georgia Institute of Technology
{qnguyen95, sfk}@gatech.edu

Abstract

Compressive signal parameterizations such as implicit neural representations and hybrid models are increasingly central to computational imaging, yet principled tools for evaluating reconstruction fidelity at finite model size remain limited when ground truth is unavailable. We introduce a framework for predicting the reconstruction error of compressive signal parameterizations, yielding non-asymptotic, signal-specific bounds that are both theoretically sound and efficiently computable without access to the ground truth signal. Specifically, we prove that when parameterization-based compression satisfies certain natural properties, the compression error at any compression level is bounded by a simple scaled difference between model predictions at different compression levels. We verify these properties for representative model families including interpolated grids, Fourier feature networks, multi-resolution hash encodings, and tensor factorizations, and show empirically that the resulting worst-case guarantees can be efficiently adapted into signal-specific error predictors that are tighter and generalizable. Across direct fitting of synthetic and natural signals, and inverse problems including radiance field and MRI reconstruction, our method closely tracks global error curves and yields informative local error heatmaps without ground-truth access. Code is available at `https://github.com/voilalab/global_error_bound`.

1 Introduction

Computational imaging seeks to recover high-fidelity signals from indirect, noisy measurements, and modern applications increasingly demand higher resolution, higher dimensionality, and dynamic content, ranging from gigavoxel volumetric MRI [1] to spatiotemporal neural radiance fields [2, 3]. These applications introduce severe bottlenecks in memory and compute. Compressive signal representations have therefore become indispensable, and a growing body of work has explored compact parameterizations for computational imaging, including sparse grids [4], basis decompositions [5, 6], coordinate-based neural networks [7–12], multi-resolution hash encodings [13], discrete mesh or point-based models [14–16], low-rank factorizations [17–19], and other hybrid models [20, 21] that combine several of these strategies.

Despite their empirical success, these models offer limited practical guidance about what is lost under compression at finite model size. Moreover, their fidelity is often assessed by empirical comparison against a known ground truth signal [12], which is not usually available in computational imaging applications. In fields such as biomedical imaging and robotic perception, quantifying reconstruction fidelity in the absence of ground truth is essential for clinical diagnostics [1, 22–24] and operational safety [25–28]. Similar needs arise in remote sensing and edge or bandwidth-limited imaging systems [29–35], where communication, memory, and compute are tightly constrained.

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

While compression loss, or rate–distortion, is a well-studied problem in information theory [36, 37], classical results and the neural compression pipelines built on them [12, 38–40] measure the size of a compressed representation in bits. The signal parameterizations now common in computational imaging are instead collections of d real-valued parameters, so bit-level results do not directly bound their error at a given d. Where theoretical approximation error bounds do exist for neural networks, they often remain confined to asymptotic or overparametrized regimes [41, 42], providing guarantees that are too coarse to inform signal-specific compression in practice. Consequently, there exist no principled tools to determine, without exhaustive search or access to ground truth, whether a given compressive real-valued parameterization is sufficiently accurate for a target application.

To address this gap, we derive finite-parameter error bounds for widely used differentiable parameterizations, including Fourier Feature Networks (FFN [11]), multi-resolution hash encodings (Instant-NGP [13]), and tensor-factorized feature grids (Geometric Algebra Planes [19]), by establishing three structural properties of their optimal error curves: monotonicity, convexity, and zero error beyond a critical parameter budget. We further complement these guarantees with a simple prediction scheme that requires fitting only a few models at nearby compression levels, translating conservative worst-case bounds into efficiently computable and empirically tight error estimates. We validate this framework across two settings: direct signal fitting, spanning synthetic signals and natural images, and imaging inverse problems, including radiance field and accelerated MRI reconstruction.

Concretely, we make the following contributions:

- We propose a framework for bounding the compression error of any signal parameterization that satisfies three natural properties: the compression error decreases monotonically with model size, is convex in model size, and is zero above some finite critical parameter budget. Under these assumptions, the compression error of a model can be rigorously bounded using only comparisons between models with nearby parameter counts, without any knowledge of the ground truth signal.
- We prove that interpolated grids as well as three representative and popular parameterizations from the computer vision literature (Fourier feature networks [11], multi-resolution hash encodings [13], and tensor factorizations [19]) satisfy these properties under explicit assumptions, using squared-error or convex envelope formulations where needed.
- Since the worst-case theoretical bounds are provably valid but not empirically tight, we introduce a simple method to convert them into practically useful error estimates and validate these across all four representations and a variety of data types and inverse problems, including fitting natural images, synthetic images and volumes, 3D objects, and 3D radiance field and both 2D and 3D MRI reconstruction.
- Although our formal theory is stated for global reconstruction error, we demonstrate empirically that the same approach also yields informative local pixel-level error estimates that align well with the true error map, enabling efficient visualization and localization of compression artifacts without access to ground truth.

2 Related Work

Compressive parameterizations. Compact signal parameterizations have found broad application in computational imaging tasks where memory and compute are tightly constrained, including accelerated MRI [1, 22–24], robotic perception [25–28], remote sensing [29–32], and edge-constrained imaging [33–35]. Recent architectures for these tasks have converged around a few families of differentiable parameterizations, including coordinate-based neural networks [7–12], trainable spatial encodings [13], and tensor-factorized models [17–19]. In this work we focus on Fourier Feature Networks (FFNs) [11], Instant-NGP [13], and Geometric Algebra Planes (GA-Planes) [19] as representative models of each type. All three models contain an MLP component, which we restrict (for the sake of analysis) to be a shallow single-hidden-layer MLP with finite width. A more detailed discussion of each model architecture is provided in Appendix A.1.

Rate-distortion theory, compression bounds, and approximation theory. Prior theory addresses neighboring but ultimately different questions. Classical rate-distortion theory [36, 37] and neural compression pipelines based on quantization or entropy coding [12, 38–40] study compression under bit budgets through distributional distortion limits or explicit coding schemes, while complementary research [43] studies the model size required for lossless neural representation of discrete signals at a

given bit precision. We instead consider representations with d real-valued parameters and bound the error for a single signal as a function of d without assuming a source distribution or specifying a coding scheme. Separately, approximation theory [41, 42, 44] and universal approximation results [45–47] for neural networks characterize expressivity, approximation rates, and the existence of exact representations as width grows. These works provide important motivation and technical ingredients, but they do not yield non-asymptotic, signal-specific error bounds for compressive parameterizations as a function of model size that are both theoretically justified and computable without access to ground truth. Our work instead studies the geometry of the optimal error curve and shows how adjacent-model differences can be converted into non-asymptotic error bounds and practically useful predictors of compression error. We provide an expanded discussion of rate-distortion theory, approximation rates, and universal approximation results in Appendices A.2–A.4.

3 Theoretical Results

We first formalize the properties that allow us to translate model differences between neighboring compression levels into rigorous error bounds. Let $x \in \mathbb{R}^n$ denote the true signal we aim to represent, such as an image or volume, and let $g_d : \mathbb{R}^d \to \mathbb{R}^n$ be a chosen compressive parameterization at compression level d/n, with learnable parameters $\theta \in \mathbb{R}^d$. For each parameter budget d, we define the optimal compression error as

$$\|x - g_d^\star\| \coloneqq \min_{\theta \in \mathbb{R}^d} \|x - g_d(\theta)\|, \tag{1}$$

where $g_d^\star$ denotes the optimal size-d representation of x under parameterization architecture g and $\|\cdot\|$ denotes the ℓ_2 norm. Directly evaluating this quantity is generally intractable: the ground-truth signal x is unobserved for many computational imaging applications, and even when x is accessible the minimization over θ is nonconvex for neural or hybrid models g. This motivates our goal of bounding $\|x - g_d^\star\|$ using only the observable difference between trained models g_{d_1} and g_{d_2} at adjacent compression levels. For notational simplicity, we state the theory using consecutive parameter budgets d and $d+1$; in practice, the parameter count of most architectures cannot be increased by exactly one, so we instead compare neighboring feasible budgets $d < d'$. The same telescoping argument applies by treating the interval $[d, d']$ as a block of compression increments. To enable this analysis, we assume the optimal error curve $d \to \|x - g_d^\star\|$ obeys three natural structural properties, which we state formally below.

Assumption 1 (Monotonicity). For all $d \geq 0$,

$$\|x - g_d^\star\| \;\geq\; \|x - g_{d+1}^\star\|. \tag{A1}$$

That is, enlarging the parameter budget cannot worsen the optimal fit.

Assumption 2 (Discrete convexity). For all $d \geq 0$,

$$\|x - g_d^\star\| - \|x - g_{d+1}^\star\| \;\geq\; \|x - g_{d+1}^\star\| - \|x - g_{d+2}^\star\|. \tag{A2}$$

This is a discrete diminishing-returns condition: the marginal improvement obtained by increasing model size is larger at small d and becomes progressively smaller as d increases.

Assumption 3 (Zero error). There exists a finite $d^\star$ such that

$$\|x - g_{d^\star}^\star\| = 0. \tag{A3}$$

In other words, at some sufficiently large model size, the model g can exactly parameterize x.

Assumption A3 is used only to anchor the telescoping argument: Theorem 10 in Appendix C.2 shows that analogous bounds hold with $d^\star$ replaced by $d^\star(\epsilon)$ satisfying $\|x - g_{d^\star(\epsilon)}^\star\| \leq \epsilon$, at the cost of an additive ϵ in the error bound. Since $d^\star(\epsilon) \leq d^\star$, this trades an ϵ error floor against a smaller multiplier, and can be strictly tighter than the exact-fit version with $d^\star$.

Together, Assumptions A1–A3 impose a simple yet powerful geometry on the optimal error curve: it decreases monotonically at a diminishing rate and converges to zero at a finite model size $d^\star$, enabling the gap between optimal models at adjacent compression levels to control the compression error relative to an unknown ground truth. We make this precise in the following sections.

3.1 Main Result

We begin with the idealized setting in which the globally optimal models $g_d^\star$ and $g_{d+1}^\star$ are available. Theorem 1 is the core of our analysis in this setting: it shows that the compression error at parameter budget d can be bounded by the difference between two neighboring optimal compressive representations, scaled by the remaining gap to the exact-fit regime. Put differently, if increasing the model size by one parameter only changes the optimal representation a little, then the model at size d is already close to the ground-truth signal.

Theorem 1 (Global Error Bound)**.** *Under Assumptions A1–A3, for any parameter budget $d \leq d^\star$, the compression error is bounded by*

$$\|x - g_d^\star\| \leq (d^\star - d)\|g_d^\star - g_{d+1}^\star\|. \tag{2}$$

Proof. Writing $\|x - g_d^\star\|$ as a telescoping sum and applying Assumption A3 yields

$$\|x - g_d^\star\| = \sum_{i=1}^{d^\star - d} \|x - g_{d+i-1}^\star\| - \|x - g_{d+i}^\star\|.$$

Assumption A2 ensures each summand is no larger than the first, so

$$\|x - g_d^\star\| \leq (d^\star - d)(\|x - g_d^\star\| - \|x - g_{d+1}^\star\|) \leq (d^\star - d)\|g_d^\star - g_{d+1}^\star\|,$$

where the last step follows from the triangle inequality $\|g_d^\star - g_{d+1}^\star\| \geq \|x - g_d^\star\| - \|x - g_{d+1}^\star\|$. □

Theorem 1 is model-agnostic and efficiently computable: it reduces the unknown error $\|x - g_d^\star\|$ relative to the ground truth to an efficiently computable quantity involving only two models at adjacent compression levels, where the scale factor $(d^\star - d)$ can be determined analytically from the architecture specification of g. Critically, this bound does not require any knowledge of the target signal x, only the difference between two model outputs and the known zero-error budget $d^\star$.

Remark 1 (Squared-error variant)**.** *For architectures where Assumption A2 holds only for the squared error, i.e. $\|x - g_d^\star\|^2 - \|x - g_{d+1}^\star\|^2 \geq \|x - g_{d+1}^\star\|^2 - \|x - g_{d+2}^\star\|^2$, an analogous telescoping argument using Assumption A1 yields*

$$\|x - g_d^\star\| \leq 2(d^\star - d)\|g_d^\star - g_{d+1}^\star\|,$$

a bound that differs from Theorem 1 only by a constant factor of 2. This variant is useful for architectures whose approximation theory is more naturally expressed in terms of squared residuals.

In practice, estimating $\|g_d^\star - g_{d+1}^\star\|$ from a single adjacent pair of models can be sensitive to training instabilities. In Appendix C.3, we provide a batched variant of Theorem 1 that averages the measured gap over a contiguous block of b compression increments (parameter counts), yielding smoother bound estimates at the cost of b additional model fits.

3.2 Handling the Optimization Gap

Theorem 1 and its batched variant are stated in terms of globally optimal models, meaning the minimization in Equation (1) is solved exactly. In practice, compressive parameterizations g are often fit by nonconvex optimization, making it difficult to ensure that the trained model at a given parameter count is globally optimal. To account for this gap between theory and practice, let $\tilde{g}_d$ denote the output of a model trained to practical convergence at size d, and define the optimization error $\eta_d \coloneqq \tilde{g}_d - g_d^\star$ as the gap between the trained model output and optimal size-d model output. Theorem 2 shows that our global error bound extends cleanly to this setting, with the bound degrading in proportion to the size and consistency of optimization errors across compression levels.

Theorem 2 (Error Bound with Optimization Gap)**.** *Under Assumptions A1–A3, for any converged model outputs $\tilde{g}_d, \tilde{g}_{d+1}$ with optimization errors η_d, η_{d+1}, the compression error is bounded by*

$$\|x - \tilde{g}_d\| \leq (d^\star - d)\|\tilde{g}_d - \tilde{g}_{d+1}\| + (d^\star - d)\|\eta_{d+1} - \eta_d\| + \|\eta_d\|. \tag{3}$$

The three terms in Theorem 2 admit clear interpretations. The first is the same neighboring-model discrepancy that appears in Theorem 1 and is directly computable. The second penalizes optimization inconsistency across adjacent compression levels: if one model is trained substantially better than its

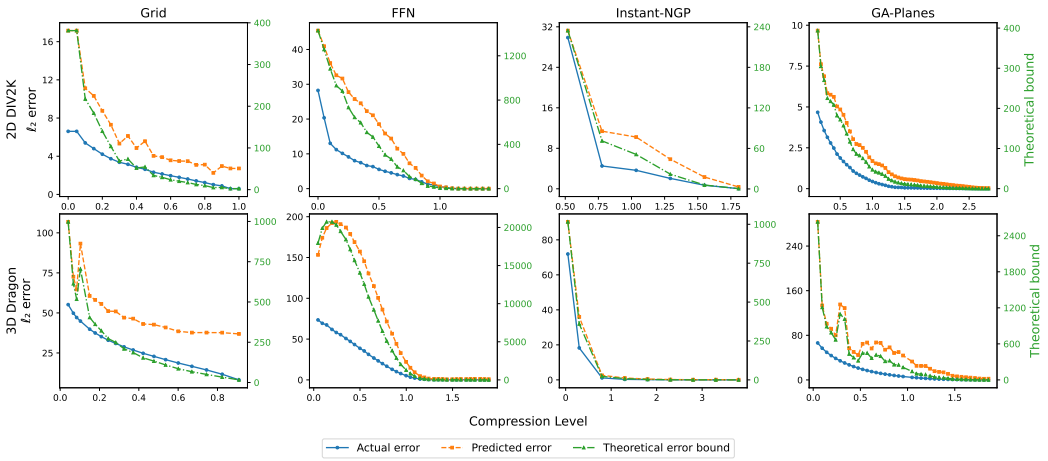

Figure 1: **Global error curves for direct fitting of signals.** We fit Grid, FFN, Instant-NGP, and GA-Planes directly to a 2D natural image from DIV2K [48] and the 3D Stanford Dragon surface volume [49]. The actual error (blue, left axis) is the ℓ_2 distance between reconstruction and ground truth, the optimization-gap-adjusted theory-derived bound (green, right axis) follows from Theorem 2 as a function of compression level, and the predicted error (orange, left axis) rescales this bound by the empirical constant c applied to the per-increment quantity $B(d)$ (Section 4.1). Although the theory-derived bounds are conservative by orders of magnitude, the calibrated predictions track the actual error across architectures, signals, and dimensionalities.

neighbor, the observed gap $\|\tilde{g}_d - \tilde{g}_{d+1}\|$ becomes a poor surrogate for $\|g^\star_d - g^\star_{d+1}\|$. The third is the irreducible optimization error at the parameter budget of interest. When $\eta_d = \eta_{d+1} = 0$, the bound in Theorem 2 reduces exactly to the bound in Theorem 1. More broadly, this result highlights that the local gap between nearby trained models remains a useful estimate of the true compression error as long as optimization behavior does not vary too erratically across adjacent compression levels.

3.3 Adaptation to Upper/Lower Bound Envelopes

For some model families, Assumption A2 cannot be readily verified for the true optimal error $\|x - g^\star_d\|$. In such cases, it suffices to replace the optimal error curve with tractable upper and lower envelopes that preserve the desired geometry. This allows us to retain the same general strategy to bound the compression error, at the cost of an additive slack term reflecting the width of the envelope. Concretely, suppose there exist functions $L, U : \mathbb{Z}_+ \to \mathbb{R}_+$ satisfying

$$L(d) \le \|x - g^\star_d\| \le U(d) \;\; \forall d, \qquad U(d^\star) = L(d^\star) = 0,$$

where the upper bound U is monotone and discretely convex. The lower envelope L need not be convex; it serves only to quantify how loose the upper envelope may be at each parameter budget.

Theorem 3 (Error Bound via Envelopes)**.** *Suppose the error curve satisfies Assumptions A1 and A3, and suppose there exist envelopes L and U satisfying the conditions above, with U monotone and discretely convex. Then the compression error is bounded by*

$$\|x - g^\star_d\| \le (d^\star - d)\big(\|g^\star_d - g^\star_{d+1}\| + U(d) - L(d)\big). \tag{4}$$

With optimization gap,

$$\|x - \tilde{g}_d\| \le (d^\star - d)\big(\|\tilde{g}_d - \tilde{g}_{d+1}\| + \|\eta_{d+1} - \eta_d\| + U(d) - L(d)\big) + \|\eta_d\|. \tag{5}$$

Theorem 3 is important conceptually because it broadens the scope of the framework in Theorem 1 by showing that discrete convexity of the true error curve is not strictly necessary. What matters is that the error curve can be bounded between surrogates with appropriate properties. The slack term $U(d) - L(d)$ quantifies the price of this relaxation: when $U = L$, the bound reduces to Theorem 1, while looser envelopes yield valid but more conservative bounds on the compression error.

3.4 Verification for Representative Architectures

The previous theorems are abstract and apply to any signal parameterization whose error curve satisfies the stated structural assumptions. We now verify that the four representations we study

satisfy the appropriate conditions, and identify which theorem governs each representation. The four models are: a Grid baseline (direct interpolated pixel/voxel parameterization), Fourier feature networks (FFN) [11], Instant-NGP [13], and 2D linear-decoder GA-Planes [19] (see the GA-Planes formulations in Appendix B.7).

Theorem 4. *Let $x \in \mathbb{R}^n$. Consider Grid, FFN and Instant-NGP (INGP) with one-hidden-layer ReLU decoders, and a 2D GA-Planes (GAP) with a linear decoder. Then:*

1. *All four models satisfy Assumptions A1 and A3, with the FFN A3 guarantee holding with high probability over its random Fourier features.*
2. *Grid satisfies the squared-error version of Assumption A2 when x has radially nonincreasing power spectrum, while GAP satisfies it unconditionally. Thus, Remark 1 applies.*
3. *Under the assumption of Siegel and Xu [42] and Achour et al. [41], FFN admits monotone, discretely convex upper bound $U(d)$ and lower bound $L(d)$ with $L(d) \le \|x - g_d^\star\| \le U(d)$. Under uniform, level-independent hashing, INGP admits such envelopes for its expected error. Hence, Theorem 3 applies.*

Proofs and additional theoretical results are provided in Appendices B and C.

4 Experiments

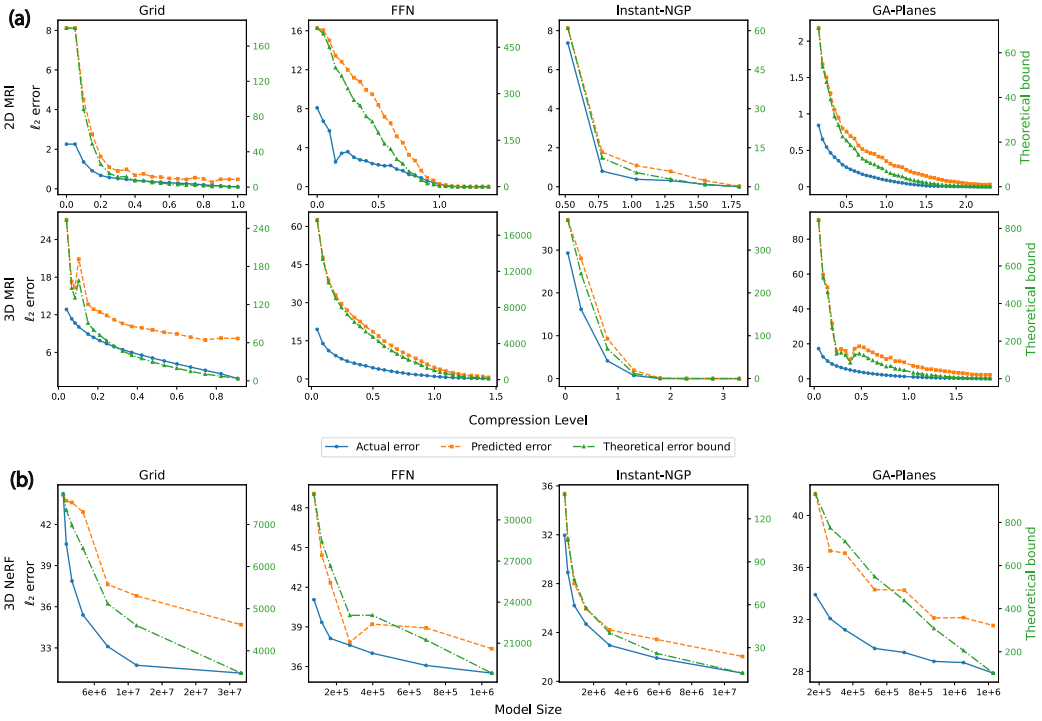

Figure 2: **Global error curves for inverse problems.** We evaluate Grid, FFN, Instant-NGP, and GA-Planes on three imaging tasks across two settings: (a) on 2D and 3D MRI reconstruction, and (b) radiance field reconstruction. MRI results are plotted against compression level d/n while NeRF results are plotted against model size, as the ground-truth signal dimension is not well-defined for radiance fields. For MRI, the actual error (blue, left axis) is the ℓ_2 distance between reconstruction and ground truth; for NeRF, it is the average ℓ_2 distance between rendered test views and the corresponding ground-truth test views. The optimization-gap-adjusted theory-derived bound (green, right axis) follows from Theorem 2 as a function of model size, and the predicted error (orange, left axis) rescales this bound by the constant c applied to the per-increment quantity $B(d)$ (Section 4.1). Despite the indirect supervision of these inverse problems, the predicted error continues to track the actual reconstruction error across compression levels and model sizes.

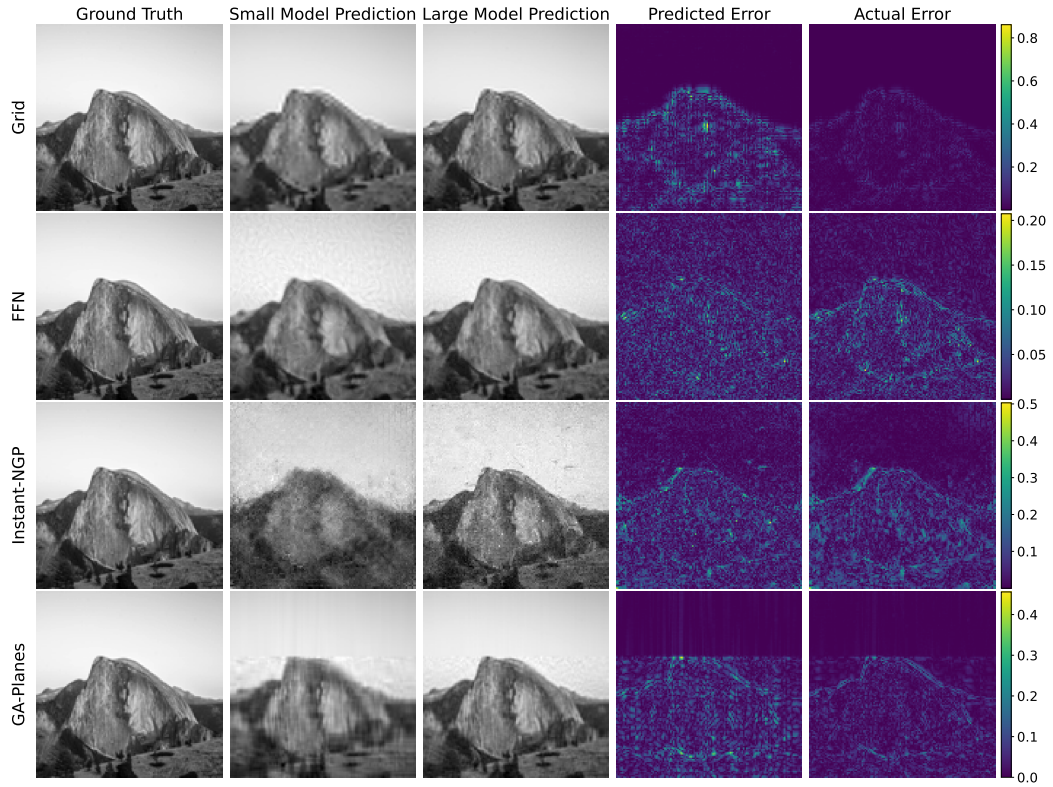

Figure 3: **Local error heatmaps for direct fitting of a DIV2K [48] natural image.** Each row corresponds to a different architecture, with small and large models at compression levels $d/n \approx 0.2$ and 0.4, respectively. Across all four representations, the predicted error heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ captures the spatial structure of the actual error heatmap $|\tilde{g}_{\text{small}} - x|$, showing that local compression artifacts can be estimated efficiently without ground-truth access.

4.1 Calibrating worst-case bounds into practical predictors

The bounds in Section 3 are provably valid and distribution-free, but conservative in practice. The factor $(d^\star - d)$ is worst-case tight under Assumptions A1–A3 alone (Theorem 8, Appendix C.1), so tighter bounds require further assumptions. If Assumption A2 is strengthened to a successive geometric decay of marginal error at rate $\rho < 1$, Theorem 9 shows that the multiplier $(d^\star - d)$ can collapse to $(1 - \rho)^{-1}$. Motivated by this refinement and the observation that the local difference between compressed reconstructions at adjacent model sizes already captures the shape of the true error curve in practice, we use the theorems as a structural template and introduce a single architecture-specific constant c that converts our theoretical bound into a practical, signal-adaptive predictor of compression error. For $d < d^\star$, define the per-increment quantity

$$B(d) \coloneqq \|\tilde{g}_d - \tilde{g}_{d+1}\| + \|\eta_{d+1} - \eta_d\| + \frac{\|\eta_d\|}{d^\star - d},$$

so that Theorem 2 reads $\|x - \tilde{g}_d\| \leq (d^\star - d)B(d)$. For each architecture, we choose the smallest c such that the predictor $cB(d)$ dominates the observed errors on a calibration signal from the ATLAS dataset [50], then hold it fixed across subsequent signals, datasets, and compression levels. Note that we calibrate separately for radiance fields using the Lego scene in NeRF-Blender [51], with the constants listed in Table 4. We quantitatively evaluate the stability of this constant c across 30 signals in Appendix D.6.

4.2 Estimating the optimization gap

Theorem 2 introduces the optimization error $\eta_d = \tilde{g}_d - g_d^\star$, the gap between the converged reconstruction and the global optimum at compression level d/n. Since $g_d^\star$ is unobservable, we estimate $\|\eta_d\|$ from restart variability. For a fixed signal, architecture, and compression level, we obtain the reconstructions $\tilde{g}_d^{(1)}, \tilde{g}_d^{(2)}, \ldots, \tilde{g}_d^{(m)}$ by training the model m times from independent random

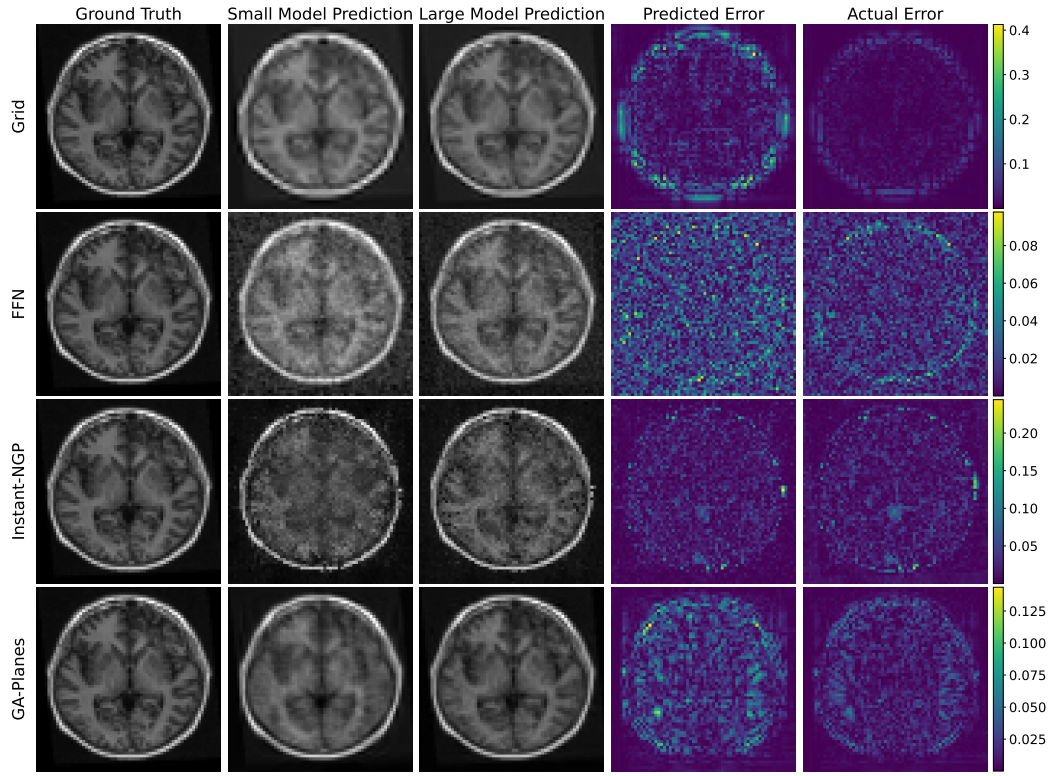

Figure 4: **Local error heatmaps for 3D MRI volume reconstruction.** Each row corresponds to a different architecture, with small and large models at compression levels $d/n \approx 0.2$ and 0.4, respectively. Each model maps 3D coordinates to signal values and is supervised indirectly through a k-space loss on a 3D volume from the ATLAS dataset [50]. A single axial slice is shown for visualization. Across all four representations, the predicted error heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ captures the spatial structure of the actual error heatmap $|\tilde{g}_{\text{small}} - x|$, indicating that reconstruction artifacts can be localized efficiently without ground-truth access, even for 3D inverse problems.

initializations with identical hyperparameters and convergence criteria, and compute the maximum pairwise ℓ_2 distance $h_m = \max_{i \neq j} \|\tilde{g}_d^{(i)} - \tilde{g}_d^{(j)}\|$. Since h_m is nondecreasing in m, we add restarts until $|h_m - h_{m-1}|$ falls below a tolerance, and use the resulting plateau value as a practical estimate of a bound on $\|\eta_d\|$ at compression level d/n.

The intuition is that variability across independently converged solutions provides an observable proxy for how far practical training may be from the ideal optimum. If at least one restart is close to $g_d^\star$, then h_m controls the optimization error of every other restart via the triangle inequality; if all runs converge to similar outputs, the estimated $\|\eta_d\|$ is small. This estimate cannot be certified as an upper bound on $\|\eta_d\|$ since multiple runs may still converge to the same suboptimal minimum, but it provides a practical, ground-truth-free diagnostic for optimization stability. We apply this correction to the global error curves in Figures 1 and 2 (Appendix D.3), which therefore report the full optimization-gap-adjusted bound of Theorem 2. In Appendix D.7, we show that the unadjusted adjacent-model predictor requires $1.99\times$–$2.01\times$ the computation time of fitting a single model, while the restart-adjusted predictor used for our global error results requires $5.09\times$–$6.78\times$.

4.3 Global error curves

Figures 1 and 2 evaluate the calibrated adjacent-model predictor across direct fitting and inverse problem settings, varying the primary capacity parameter of each architecture while holding other hyperparameters fixed. Across all settings, the predicted error captures the overall shape and scale of the true error curves across architectures, datasets, and tasks. The theoretical bounds remain valid but conservative, whereas the calibrated predictions are substantially tighter and preserve the qualitative decay of the true reconstruction error. The actual error curves are also approximately convex and monotonically decreasing up to optimization variability, providing empirical support for the structural

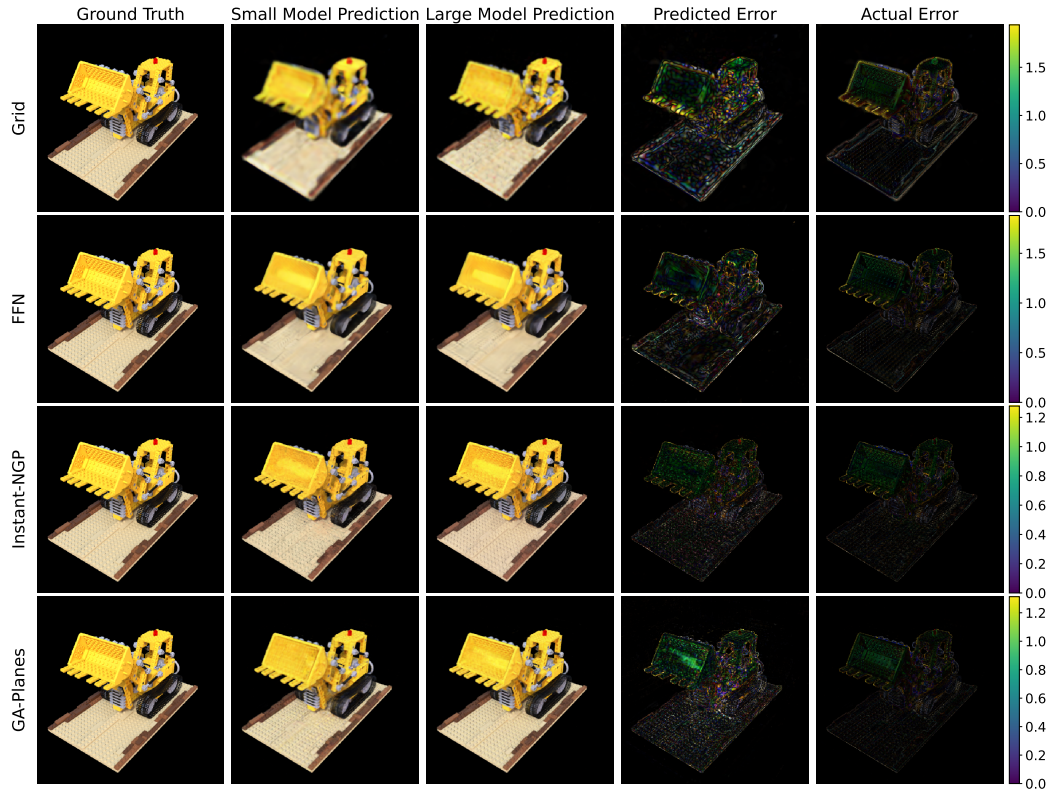

Figure 5: **Local error heatmaps for 3D radiance field reconstruction.** Each row corresponds to a different architecture, with small and large models having 2.5×10^5 and 1.6×10^6 parameters, respectively. Each model represents a 3D radiance field supervised through a photometric loss on training views from the Lego scene in NeRF-Blender [51]; a single test viewpoint is shown. Across all four representations, the predicted error heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ captures the spatial structure of the true error heatmap, indicating that reconstruction artifacts can be localized efficiently at novel viewpoints without access to the ground-truth test view.

assumptions underlying our theory. These results suggest that adjacent-model differences provide a stable global proxy for reconstruction error in both direct fitting and inverse-problem settings, despite the latter involving indirect supervision and nontrivial forward models, and that the looseness of the worst-case bound is likely architecture-dependent rather than signal-dependent or task-dependent. In Appendix D.6, we repeat the full model capacity sweep on 30 signals and find that the mean predicted-to-actual error ratio lies between 1.91 and 2.54, confirming the approximate factor-of-two conservatism visible in Figures 1 and 2. Signal-specific error prediction also motivates ground-truth-free model-size selection under a prescribed fidelity target; Appendix D.5 provides an illustrative MRI example in which statistical (dataset-based) error estimates can underestimate error on new or challenging signals, while our calibrated error estimates adapt and remain conservative across signals.

We note that the NeRF experiments extend beyond our formal shallow-decoder theory in two ways: the decoders are multi-layer nonlinear MLPs (rather than the single-hidden-layer decoders covered by our proofs), and they take additional view-direction inputs to model view-dependent radiance, as is standard in NeRF [51]. Full architectural details are in Appendix D.2. These experiments are therefore intended as empirical stress tests of whether the adjacent-model predictor remains useful for more challenging tasks like radiance field reconstruction. Additional details on test signals and global error curves for synthetic signals are provided in Appendices D.1 and D.3.

4.4 Local error heatmaps

Although our theory is stated for global ℓ_2 error, the same adjacent-model principle can be applied pointwise to obtain spatially resolved error estimates. Given a small and large model from the same representation family, we compute the predicted error heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ using the constant c

from Section 4.1. This estimate requires only the two trained reconstructions and no access to ground truth. For evaluation only, we compare it against the actual error heatmap $|\tilde{g}_{\text{small}} - x|$.

We evaluate local error prediction for direct fitting on 2D natural images from DIV2K [48] (Figure 3), 3D MRI volume reconstruction from k-space measurements using the ATLAS dataset [50] (Figure 4), and 3D radiance field reconstruction of the Lego scene [51] (Figure 5). Across all architectures and tasks, regions of high predicted error consistently coincide with high actual error, while low-error regions are correctly identified. We emphasize that these heatmaps carry no formal theoretical guarantee, as our bounds are stated for the global ℓ_2 norm, not pointwise or region-wise error. However, their consistency suggests that adjacent-model differences are informative not only in aggregate, but also about the spatial locations where a compressed reconstruction is likely to be unreliable. Further implementation details, additional results for other tasks, and quantitative metrics for the error heatmaps may be found in Appendices D.2 and D.4.

5 Discussion

In this work, we present, to our knowledge, the first framework for non-asymptotic, signal-specific error bounds for compressive signal parameterizations that are computable without access to ground truth. Under three structural conditions on the optimal error curve, which we show to hold for representative signal parameterizations, we prove that the reconstruction error at a given model size is bounded by the difference between compressed reconstructions at adjacent model sizes. We complement our theory with a simple scalar adaptation that yields predicted error curves closely tracking the true error across diverse applications. We further find empirically that the same local-comparison principle produces informative pixel-level error heatmaps, despite our theory being stated only for global error.

Limitations and future work. Our theoretical guarantees rely on structural assumptions about the optimal error curve that we verify only for the model families studied here, each using shallow, scalar-output MLP decoders. Extending these results to deeper, vector-output decoders and a broader set of architectures and input encodings is an important direction for future work. Our scalar adaptation from worst-case bounds to practical predictions is estimated via a simple heuristic calibration; a more principled calibration procedure could yield tighter error estimates. Finally, our theoretical results bound global reconstruction error; the local heatmaps remain empirical error estimates rather than proven error bounds. Extending our theory to local seminorms or region-of-interest metrics would provide theoretical grounding for the spatially resolved predictions that our experiments suggest are informative, and would strengthen the connection to safety-critical applications where localized compression artifacts can matter more than average error.

References

[1] Frank Ong, Xucheng Zhu, Joseph Y Cheng, Kevin M Johnson, Peder EZ Larson, Shreyas S Vasanawala, and Michael Lustig. Extreme MRI: Large-scale volumetric dynamic imaging from continuous non-gated acquisitions. *Magnetic Resonance in Medicine*, 84(4):1763–1780, 2020.

[2] Chenyue Jiao, Chongke Bi, and Lu Yang. FFEINR: Flow Feature-Enhanced Implicit Neural Representation for Spatiotemporal Super-Resolution. *Journal of Visualization*, 27(2):273–289, 2024.

[3] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, 2021.

[4] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[5] Richard H Bartels, John C Beatty, and Brian A Barsky. *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann, 1995.

[6] S Allen Broughton and Kurt Bryan. *Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing*. John Wiley & Sons, 2018.

[7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning Continuous Image Representation with Local Implicit Image Function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021.

[8] David B. Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. BACON: Band-limited Coordinate Networks for Multiscale Scene Representation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16231–16241, 2022. doi: 10.1109/CVPR52688.2022.01577.

[9] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. WIRE: Wavelet Implicit Neural Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18507–18516, 2023.

[10] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:7537–7547, 2020.

[12] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. Implicit Neural Representations for Image Compression. In *European Conference on Computer Vision*, pages 74–91. Springer, 2022.

[13] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics (TOG)*, 41 (4):1–15, 2022.

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`.

[15] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. MeshGPT: Generating Triangle Meshes with Decoder-Only Transformers. *arXiv preprint arXiv:2311.15475*, 2023.

[16] Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous PDE Dynamics Forecasting with Implicit Neural Representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=B73niNjbPs`.

[17] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial Radiance Fields. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, page 333–350, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19823-6. doi: 10.1007/978-3-031-19824-3_20. URL `https://doi.org/10.1007/978-3-031-19824-3_20`.

[18] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] Irmak Sivgin, Sara Fridovich-Keil, Gordon Wetzstein, and Mert Pilanci. Geometric Algebra Planes: Convex Implicit Neural Volumes. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=yHpWI6a2xT`.

[20] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Dictionary Fields: Learning a Neural Basis Decomposition. *ACM Trans. Graph.*, 42(4), July 2023. ISSN 0730-0301. doi: 10.1145/3592135. URL `https://doi.org/10.1145/3592135`.

[21] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.*, 40(4), 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459863. URL https://doi.org/10.1145/3450626.3459863.

[22] Reinhard Heckel, Mathews Jacob, Akshay Chaudhari, Or Perlman, and Efrat Shimron. Deep Learning for Accelerated and Robust MRI Reconstruction: a Review, 2024. URL https://arxiv.org/abs/2404.15692.

[23] Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi: https://doi.org/10.1002/mrm.21391. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21391.

[24] Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit Neural Representation in Medical Imaging: A Comparative Survey, 2023. URL https://arxiv.org/abs/2307.16142.

[25] Guangming Wang, Lei Pan, Songyou Peng, Shaohui Liu, Chenfeng Xu, Yanzi Miao, Wei Zhan, Masayoshi Tomizuka, Marc Pollefeys, and Hesheng Wang. NeRFs in Robotics: A Survey, 2025. URL https://arxiv.org/abs/2405.01333.

[26] Bernardo Araújo, João F. Teixeira, Joaquim Fonseca, Ricardo Cerqueira, and Sofia C. Beco. The Road to Safety: A Review of Uncertainty and Applications to Autonomous Driving Perception. *Entropy*, 26(8), 2024. ISSN 1099-4300. doi: 10.3390/e26080634. URL https://www.mdpi.com/1099-4300/26/8/634.

[27] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. NICE-SLAM: Neural Implicit Scalable Encoding for SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[28] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew Davison. iMAP: Implicit Mapping and Positioning in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[29] KS Gunasheela and HS Prasantha. Satellite Image Compression-Detailed Survey of the Algorithms. In *Proceedings of International Conference on Cognition and Recognition: ICCR 2016*, pages 187–198. Springer, 2018.

[30] Carlos Gomes, Isabelle Wittmann, Damien Robert, Johannes Jakubik, Tim Reichelt, Stefano Maurogiovanni, Rikard Vinge, Jonas Hurst, Erik Scheurer, Rocco Sedona, Thomas Brunschwiler, Stefan Kesselheim, Matej Batič, Philip Stier, Jan Dirk Wegner, Gabriele Cavallaro, Edzer Pebesma, Michael Marszalek, Miguel A. Belenguer-Plomer, Kennedy Adriko, Paolo Fraccaro, Romeo Kienzler, Rania Briq, Sabrina Benassou, Michele Lazzarini, and Conrad M. Albrecht. Lossy Neural Compression for Geospatial Analytics: A review. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):97–135, September 2025. ISSN 2473-2397. doi: 10.1109/mgrs.2025.3546527. URL http://dx.doi.org/10.1109/MGRS.2025.3546527.

[31] Mireille Louys, Jean-Luc Starck, Simona Mei, François Bonnarel, and Fionn Murtagh. Astronomical image compression. *Astronomy and Astrophysics Supplement Series*, 136(3):579–590, 1999.

[32] AR Offringa. Compression of interferometric radio-astronomical data. *Astronomy & Astrophysics*, 595:A99, 2016.

[33] Dat Ngo, Hyun-Cheol Park, and Bongsoon Kang. Edge Intelligence: A Review of Deep Neural Network Inference in Resource-Limited Environments. *Electronics*, 14(12), 2025. ISSN 2079-9292. doi: 10.3390/electronics14122495. URL https://www.mdpi.com/2079-9292/14/12/2495.

[34] Zengshan Yin, Changhao Wu, Chongbin Guo, Yuanchun Li, Mengwei Xu, Weiwei Gao, and Chuanxiu Chi. A Comprehensive Survey of Orbital Edge Computing: Systems, Applications, and Algorithms. *Chinese Journal of Aeronautics*, 38(7):103316, July 2025. ISSN 1000-9361. doi: 10.1016/j.cja.2024.11.026. URL http://dx.doi.org/10.1016/j.cja.2024.11.026.

[35] R. Del Prete, P. K. Thind, A. Mazzeo, M. Whitley, L. Papa, N. Longépé, and G. Meoni. Optimizing deep learning models for on-orbit deployment through neural architecture search. *Scientific Reports*, 15(1), 2025. doi: 10.1038/s41598-025-21467-8. URL https://doi.org/10.1038/s41598-025-21467-8.

[36] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006.

[37] C. E. Shannon. Probability of Error for Optimal Codes in a Gaussian Channel. *Bell System Technical Journal*, 38:611–656, 1959. doi: 10.1002/j.1538-7305.1959.tb03905.x.

[38] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. NeRV: Neural Representations for Videos. In *NeurIPS*, 2021.

[39] G. Dai, R. Zhang, Q. Wuwu, et al. Implicit Neural Image Field for Biological Microscopy Image Compression. *Nature Computational Science*, 5:1041–1050, 2025. doi: 10.1038/s43588-025-00889-4. URL https://doi.org/10.1038/s43588-025-00889-4.

[40] Zongyu Guo, Gergely Flamich, Jiajun He, Zhibo Chen, and José Miguel Hernández-Lobato. Compression with Bayesian Implicit Neural Representations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 1938–1956. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/060b2af0081a460f7f466f7f174d9052-Paper-Conference.pdf.

[41] El Mehdi Achour, Armand Foucault, Sébastien Gerchinovitz, and Francois Malgouyres. A general approximation lower bound in L^p norm, with applications to feed-forward neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=tfkeJG9yAX.

[42] Jonathan W. Siegel and Jinchao Xu. Approximation Rates for Neural Networks with General Activation Functions. *Neural Networks*, 128:313–321, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.05.019. URL https://www.sciencedirect.com/science/article/pii/S0893608020301891.

[43] Woo Kyoung Han, Byeonghun Lee, Hyunmin Cho, Sunghoon Im, and Kyong Hwan Jin. Towards Lossless Implicit Neural Representation via Bit Plane Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[44] Andrew R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

[45] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

[46] Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum Width for Universal Approximation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O-XJwyoIF-k.

[47] Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using ReLU networks on compact domain. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dpDw5U04SU.

[48] Eirikur Agustsson and Radu Timofte. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[49] Brian Curless and Marc Levoy. A Volumetric Method for Building Complex Models from Range Images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

[50] Sook-Lei Liew, Bethany Lo, Miranda R. Donnelly, Artemis Zavaliangos-Petropulu, Jessica N. Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P. Simon, Julia M. Juliano, Anisha Suri, Tyler Ard, Nerisa Banaj, Michael R. Borich, Lara A. Boyd, Amy Brodtmann, Cathrin M. Buetefisch, Lei Cao, Jessica M. Cassidy, Valentina Ciullo, Adriana B. Conforto, Steven C. Cramer, Rosalia Dacosta-Aguayo, Ezequiel de la Rosa, Martin Domin, Adrienne N. Dula, Wuwei Feng, Alexandre R. Franco, Fatemeh Geranmayeh, Alexandre Gramfort, Chris M. Gregory, Colleen A. Hanlon, Brenton G. Hordacre, Steven A. Kautz, Mohamed Salah Khlif, Hosung Kim, Jan S. Kirschke, Jingchun Liu, Martin Lotze, Bradley J. MacIntosh, Maria Mataró, Feroze B. Mohamed, Jan E. Nordvik, Gilsoon Park, Amy Pienta, Fabrizio Piras, Shane M. Redman, Kate P. Revill, Mauricio Reyes, Andrew D. Robertson, Na Jin Seo, Surjo R. Soekadar, Gianfranco Spalletta, Alison Sweet, Maria Telenczuk, Gregory Thielman, Lars T. Westlye, Carolee J. Winstein, George F. Wittenberg, Kristin A. Wong, and Chunshui Yu. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *medRxiv*, 2021. doi: 10.1101/2021.12.09.21267554. URL `https://www.medrxiv.org/content/early/2021/12/11/2021.12.09.21267554`.

[51] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

[52] Namhoon Kim and Sara Fridovich-Keil. Grids Often Outperform Implicit Neural Representation at Compressing Dense Signals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=OZljvntsto`.

[53] A. van der Schaaf and J. H. van Hateren. Modelling the Power Spectra of Natural Images: Statistics and Information. *Vision Research*, 36(17):2759–2770, 1996.

[54] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.

[55] Yueyang Xu, Ashish Raj, and Jonathan D. Victor. Systematic Differences Between Perceptually Relevant Image Statistics of Brain MRI and Natural Images. *Frontiers in Neuroinformatics*, Volume 13 - 2019, 2019. ISSN 1662-5196. doi: 10.3389/fninf.2019.00046. URL `https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2019.00046`.

[56] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL `http://jmlr.org/papers/v20/17-612.html`.

[57] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011. URL `https://arxiv.org/abs/1011.3027`.

[58] Matthias Teschner, Bruno Heidelberger, Matthias Müller, Danat Pomerantes, and Markus Gross. Optimized Spatial Hashing for Collision Detection of Deformable Objects. In *Proc. VMV*, pages 47–54, 2003.

[59] Nelson Max. Optical Models for Direct Volume Rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995.

[60] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018.

Appendices

Norm notation. Throughout the appendices, we use L^p to denote the function-space norms on continuous domains, consistent with the approximation theory literature [41, 42, 44], whereas ℓ_p denotes the finite-dimensional norm on sampled signals used in the main text. On the finite sample grid used in our experiments, the L^2 and ℓ_2 norms coincide up to a fixed normalization, so the bounds we adopt translate directly to the ℓ_2 norms used in the main text.

A Related Work

We provide here an expanded discussion of the four lines of prior work summarized in Section 2.

A.1 Representative approaches to compressive parameterizations

Compressive, continuous signal representations have found broad application in computational imaging tasks where memory and compute are tightly constrained, including accelerated MRI [1, 22–24], robotic perception [25–28], remote sensing [29–32], and edge-constrained imaging [33–35]. Recent architectures for these tasks have converged around three design paradigms: coordinate-based neural networks [7–12], trainable spatial encodings [13], and tensor-factorized models [17–19]. Fourier Feature Networks (FFNs) [11] preprocess input coordinates with a sinusoidal feature lifting before passing them to a ReLU MLP, transforming the effective neural tangent kernel into a stationary kernel with tunable bandwidth and improving the learning of high-frequency structures. Instant-NGP [13] stores learnable features in a multi-resolution hash table indexed by input coordinates, and decodes the resulting concatenated multi-scale features with a lightweight MLP. This multi-resolution construction helps the decoder disambiguate hash collisions, enabling high fidelity within a compact parameter budget. Geometric Algebra Planes (GA-Planes) [19] combines line, plane, and volume features with a neural decoder, and admits convex and semi-convex optimization formulations that in the 2D setting are provably equivalent to a (masked) low-rank plus low-resolution matrix factorization. Together, these models span three qualitatively different compression mechanisms—spectral lifting, multi-resolution feature sharing, and structured low-dimensional factorization—and approximately sample the current Pareto frontier of model size, expressiveness, and optimizability [19]. We therefore focus our theoretical and experimental analysis on these three models as well as an interpolated Grid baseline [52].

A.2 Rate-distortion theory and compression bounds

Classical rate–distortion theory characterizes the fundamental tradeoff between compression rate and reconstruction fidelity for stochastic sources, providing a lower bound on the distortion achievable at any given bit budget [36, 37]. A growing line of work has brought rate–distortion ideas into implicit neural representation-based compression pipelines for images, volumes, and video, typically by combining trained neural fields with quantization, entropy coding, or related coding mechanisms [12, 38–40]. A complementary direction studies the model size required for lossless neural representation: Han et al. [43] derive explicit upper bounds on the number of parameters needed for an INR to achieve a given bit precision on a discrete signal, showing that this threshold grows exponentially with bit depth. These works make important progress toward INR-based compression, but their primary emphasis remains codec or representation design, and evaluation is largely empirical rather than oriented around signal-specific theoretical guarantees on representation error.

Our setting differs from this literature in how representation size is measured. Rate–distortion theory and neural codecs measure size in bits, while we consider real-valued parameter vectors $\theta \in \mathbb{R}^d$ and measure size by parameter count d. These two notions of size are not interchangeable: bit-level guarantees do not characterize the error achievable by a fixed architecture at a given d and accordingly, parameter count alone does not specify a bit rate without additional choices of numerical precision and coding. Moreover, classical rate–distortion theory assumes a known or estimable probability distribution over sources [36], whereas our goal is to bound the representation error of a single signal at finite model size without distributional assumptions. Our results therefore complement bit-level analyses: a fitted parameterization may subsequently be quantized and entropy-coded, a step our framework does not model. To our knowledge, no prior work provides non-asymptotic, signal-specific error bounds for compressive representations as a function of model size, that are both

theoretically justified and computable without access to ground truth. Our framework addresses this gap by replacing distributional assumptions with three structural properties of the optimal error curve, which we verify for each architecture we study.

A.3 Approximation rates for neural networks

A substantial body of theoretical work characterizes how accurately neural networks of fixed size can approximate functions drawn from classical function spaces. Early results of Barron [44] established dimension-independent L^2 rates for two-layer networks approximating functions with bounded spectral norm. Siegel and Xu [42] generalized these results to broader activation families, providing approximation rates for polynomially decaying non-sigmoidal activations and any bounded integrable activations, and characterized how the error scales with network width and depth. Achour et al. [41] complemented these upper bounds with tight lower bounds for piecewise-polynomial feed-forward networks on Hölder balls and multivariate monotone functions. A structural implication shared by both works is that the approximation error of an MLP is bounded above by a convex, decreasing function of the number of parameters, a key property that we leverage in our main theorem. In particular, we adopt the upper bound of Siegel and Xu [42] and the lower bound of Achour et al. [41], adapted to account for the Fourier feature embedding of FFNs, as the convex envelope for the error curve in our proof of Theorem 4.

A.4 Universal approximation for neural networks

The classical Universal Approximation Theorem [45] guarantees that a sufficiently wide single-hidden-layer MLP can approximate any continuous function on a compact domain, but provides no constructive bound on the width required. Park et al. [46] resolved this question by identifying $\max\{d_x + 1, d_y\}$ as the exact minimum-width threshold for universal approximation of ReLU networks on $\mathbb{R}^{d_x}$, ruling out any strictly smaller width and establishing this threshold as a fundamental architectural limit. Here d_x denotes the input dimension and d_y the output dimension (number of scalar signal channels predicted by the decoder). Kim et al. [47] further refined this guarantee for the practically relevant compact-domain setting, proving that the minimum width for L^p approximation of L^p functions on $[0, 1]^{d_x}$ drops to $\max\{d_x, d_y, 2\}$ for ReLU-like activations, including ReLU, GELU, and Softplus. These results are the key ingredient in our proof that all three compressive representations with MLP decoders we consider satisfy Assumption A3: the existence of a finite model size at which the optimal representation error vanishes. They are invoked most directly in the proof for FFNs, where the coordinate domain is normalized to $[0, 1]^k$ and the effective decoder is a ReLU MLP composed with a Fourier feature embedding.

B Proofs of Theorems

Notation for proofs. Proofs follow the same notational conventions as the main text. For a fixed parameter budget d, let $g_d^\star$ denote the globally optimal representation, $\tilde{g}_d$ denote the output of a model trained to practical convergence, and $\eta_d = \tilde{g}_d - g_d^\star$ be the optimization error.

B.1 Proof of Remark 1

Proof. Suppose Assumption A2 holds only for the squared error, i.e., $\|x - g_d^\star\|^2 - \|x - g_{d+1}^\star\|^2 \geq \|x - g_{d+1}^\star\|^2 - \|x - g_{d+2}^\star\|^2$ for all $d \geq 0$. Writing $\|x - g_d^\star\|^2$ as a telescoping sum and applying Assumption A3 yields

$$\|x - g_d^\star\|^2 = \sum_{i=1}^{d^\star - d} \|x - g_{d+i-1}^\star\|^2 - \|x - g_{d+i}^\star\|^2.$$

It follows from the assumption that each summand is no larger than the first, so

$$\|x - g_d^\star\|^2 \leq (d^\star - d)(\|x - g_d^\star\|^2 - \|x - g_{d+1}^\star\|^2).$$

Factoring the difference of squares and applying the triangle inequality $\|x - g_d^\star\| - \|x - g_{d+1}^\star\| \le \|g_d^\star - g_{d+1}^\star\|$, we obtain

$$\begin{aligned}\|x - g_d^\star\|^2 - \|x - g_{d+1}^\star\|^2 &= (\|x - g_d^\star\| - \|x - g_{d+1}^\star\|)(\|x - g_d^\star\| + \|x - g_{d+1}^\star\|) \\ &\le \|g_d^\star - g_{d+1}^\star\|(\|x - g_d^\star\| + \|x - g_{d+1}^\star\|) \\ &\le \|g_d^\star - g_{d+1}^\star\|(\|x - g_d^\star\| + \|x - g_d^\star\|) = 2\|g_d^\star - g_{d+1}^\star\|\|x - g_d^\star\|,\end{aligned}$$

where the third line uses Assumption A1. Combining the two inequalities gives

$$\|x - g_d^\star\|^2 \le 2(d^\star - d)\|g_d^\star - g_{d+1}^\star\|\|x - g_d^\star\|.$$

If $\|x - g_d^\star\| = 0$, the bound is trivially true; otherwise, dividing both sides by $\|x - g_d^\star\|$ yields

$$\|x - g_d^\star\| \le 2(d^\star - d)\|g_d^\star - g_{d+1}^\star\|.$$

□

B.2 Proof of Theorem 2

Proof. By the triangle inequality and the identity $\tilde{g}_d = g_d^\star + \eta_d$, we have

$$\|x - \tilde{g}_d\| = \|x - g_d^\star - \eta_d\| \le \|x - g_d^\star\| + \|\eta_d\|.$$

From Theorem 1, we have

$$\|x - g_d^\star\| \le (d^\star - d)\|g_d^\star - g_{d+1}^\star\|.$$

Therefore,

$$\|x - \tilde{g}_d\| \le (d^\star - d)\|g_d^\star - g_{d+1}^\star\| + \|\eta_d\|.$$

Substituting $g_d^\star = \tilde{g}_d - \eta_d$ and $g_{d+1}^\star = \tilde{g}_{d+1} - \eta_{d+1}$, we obtain

$$\begin{aligned}\|x - \tilde{g}_d\| &\le (d^\star - d)\|\tilde{g}_d - \eta_d - (\tilde{g}_{d+1} - \eta_{d+1})\| + \|\eta_d\| \\ &\le (d^\star - d)\Big(\|\tilde{g}_d - \tilde{g}_{d+1}\| + \|\eta_{d+1} - \eta_d\|\Big) + \|\eta_d\| \\ &= (d^\star - d)\|\tilde{g}_d - \tilde{g}_{d+1}\| + (d^\star - d)\|\eta_{d+1} - \eta_d\| + \|\eta_d\|,\end{aligned}$$

which is exactly the claimed bound. □

B.3 Proof of Theorem 3

Proof. Define the slack

$$\xi_d := U(d) - \|x - g_d^\star\|.$$

Since $L(d) \le \|x - g_d^\star\| \le U(d)$ by assumption, it follows that

$$0 \le \xi_d \le U(d) - L(d).$$

Since U is monotone, discretely convex, and satisfies $U(d^\star) = 0$, the same telescoping and discrete-convexity argument used in the proof of Theorem 1 applies directly to U. In particular,

$$U(d) = \sum_{i=1}^{d^\star - d} \big(U(d + i - 1) - U(d + i)\big) \le (d^\star - d)\big(U(d) - U(d + 1)\big).$$

It therefore remains to bound the one-step decrement $U(d) - U(d+1)$. Since $\|x - g_{d+1}^\star\| \le U(d+1)$, we have

$$\begin{aligned}U(d) - U(d + 1) &\le U(d) - \|x - g_{d+1}^\star\| \\ &= \big(U(d) - \|x - g_d^\star\|\big) + \big(\|x - g_d^\star\| - \|x - g_{d+1}^\star\|\big) \\ &= \xi_d + \|x - g_d^\star\| - \|x - g_{d+1}^\star\| \\ &\le \xi_d + \|g_d^\star - g_{d+1}^\star\|,\end{aligned}$$

where the last step follows from the triangle inequality. Substituting this bound into the previous inequality and using $\xi_d \le U(d) - L(d)$ yields

$$U(d) \le (d^\star - d)\big(\|g_d^\star - g_{d+1}^\star\| + U(d) - L(d)\big).$$

Since $\|x - g_d^\star\| \le U(d)$, we can conclude that

$$\|x - g_d^\star\| \le (d^\star - d)\big(\|g_d^\star - g_{d+1}^\star\| + U(d) - L(d)\big).$$

We now incorporate the optimization gap. By the triangle inequality and the identity $\tilde{g}_d = g_d^\star + \eta_d$,

$$\|x - \tilde{g}_d\| \le \|x - g_d^\star\| + \|\eta_d\|.$$

Applying the bound just established for $\|x - g_d^\star\|$ and substituting

$$g_d^\star = \tilde{g}_d - \eta_d, \qquad g_{d+1}^\star = \tilde{g}_{d+1} - \eta_{d+1},$$

we obtain

$$\begin{aligned}\|g_d^\star - g_{d+1}^\star\| &= \|(\tilde{g}_d - \tilde{g}_{d+1}) - (\eta_d - \eta_{d+1})\| \\ &\le \|\tilde{g}_d - \tilde{g}_{d+1}\| + \|\eta_{d+1} - \eta_d\|,\end{aligned}$$

exactly as in the proof of Theorem 2. Combining the preceding inequality gives

$$\|x - \tilde{g}_d\| \le (d^\star - d)\big(\|\tilde{g}_d - \tilde{g}_{d+1}\| + \|\eta_{d+1} - \eta_d\| + U(d) - L(d)\big) + \|\eta_d\|,$$

which proves the result. □

B.4 Proof of Theorem 4: Grid

We prove the Grid representation case by indexing the model family by its spatial resolution rather than by raw parameter count. Let the target signal x be sampled on an N^D lattice (which can be done losslessly for any bandlimited continuous signal, for large enough N), and let a Grid model of per-axis resolution r use $d_r = r^D$ parameters, one value at each coarse lattice point. We assume ideal sinc interpolation from the coarse r^D lattice to the fine N^D lattice. Under this interpolation, the set of signals representable by a resolution-r grid is exactly the subspace $\mathcal{S}_r$ of discrete signals whose Fourier coefficients are supported on the hypercube $\{\mathbf{k} : \|\mathbf{k}\|_\infty \le \lfloor r/2 \rfloor\}$. Consequently, the optimal Grid reconstruction $g_{d_r}^\star$ is the orthogonal projection of x onto $\mathcal{S}_r$, and the optimal squared error decomposes cleanly via Parseval's theorem:

$$\|x - g_{d_r}^\star\|^2 = \sum_{\|\mathbf{k}\|_\infty > \lfloor r/2 \rfloor} |X[\mathbf{k}]|^2,$$

where $X[\mathbf{k}]$ denotes the D-dimensional discrete Fourier transform of x. This spectral characterization of the optimal error is the foundation for all three lemmas below.

Lemma G.1 (Grid satisfies Assumption A1)**.** *For each per-axis resolution $r \ge 0$,*

$$\|x - g_{d_r}^\star\| \ge \|x - g_{d_{r+1}}^\star\|.$$

Proof. Since $\lfloor (r+1)/2 \rfloor \ge \lfloor r/2 \rfloor$, the bandlimited subspace $\mathcal{S}_r$ is contained in $\mathcal{S}_{r+1}$. The set of frequencies excluded at resolution $r+1$ is therefore a subset of those excluded at resolution r, so

$$\|x - g_{d_r}^\star\|^2 - \|x - g_{d_{r+1}}^\star\|^2 = \sum_{\lfloor r/2 \rfloor < \|\mathbf{k}\|_\infty \le \lfloor (r+1)/2 \rfloor} |X[\mathbf{k}]|^2 \ge 0.$$

Since the square root function is monotone increasing, the inequality $\|x - g_{d_r}^\star\|^2 \ge \|x - g_{d_{r+1}}^\star\|^2$ implies $\|x - g_{d_r}^\star\| \ge \|x - g_{d_{r+1}}^\star\|$. □

Lemma G.2 (Grid satisfies Assumption A2 for squared error, under radial spectral decay)**.** *Assume in addition that the target signal x has a radially nonincreasing power spectrum, meaning $|X[\mathbf{k}]|^2 \ge |X[\mathbf{k}']|^2$ whenever $\|\mathbf{k}'\|_\infty \ge \|\mathbf{k}\|_\infty$. Then for each per-axis resolution $r \ge 0$, the optimal square error satisfies the discrete diminishing-returns condition:*

$$\|x - g_{d_r}^\star\|^2 - \|x - g_{d_{r+1}}^\star\|^2 \ge \|x - g_{d_{r+1}}^\star\|^2 - \|x - g_{d_{r+2}}^\star\|^2.$$

Proof. From the projection characterization above, the decrease in optimal squared error from resolution r to resolution $r+1$ is exactly the Fourier energy contained in the shell $\lfloor r/2 \rfloor < \|\mathbf{k}\|_\infty \leq \lfloor (r+1)/2 \rfloor$, or equivalently,

$$\|x - g^\star_{d_r}\|^2 - \|x - g^\star_{d_{r+1}}\|^2 = \sum_{\lfloor r/2 \rfloor < \|\mathbf{k}\|_\infty \leq \lfloor (r+1)/2 \rfloor} |X[\mathbf{k}]|^2.$$

Likewise, the decrease from resolution $r+1$ to resolution $r+2$ is exactly the energy in the next shell $\lfloor (r+1)/2 \rfloor < \|\mathbf{k}\|_\infty \leq \lfloor (r+2)/2 \rfloor$, or equivalently,

$$\|x - g^\star_{d_{r+1}}\|^2 - \|x - g^\star_{d_{r+2}}\|^2 = \sum_{\lfloor (r+1)/2 \rfloor < \|\mathbf{k}\|_\infty \leq \lfloor (r+2)/2 \rfloor} |X[\mathbf{k}]|^2.$$

Under the radial spectral decay assumption on $|X[\mathbf{k}]|^2$, the frequencies in the earlier shell carry at least as much power as frequencies in the later shell, so the first shell energy dominates the second. Hence, the one-step improvement in squared error is nonincreasing with resolution, which is exactly discrete convexity for the squared residual. □

Remark 2. *The radial spectral decay condition is a natural regularity assumption that is broadly satisfied by many targets of interest in computational imaging, whose power spectral densities typically follow a power-law or exponential decay with frequency [23, 53–55]. Under this common low-pass structure, the energy captured by successively added frequency shells decreases with resolution, so the squared-error discrete convexity established in Lemma G.2 matches the expected approximation behavior of Grid representations with sinc interpolation.*

Lemma G.3 (Grid satisfies Assumption A3)**.** *There exists a finite per-axis resolution $r^\star$ such that $\|x - g^\star_{d_{r^\star}}\| = 0$.*

Proof. At full resolution $r^\star = N$, each grid parameter can be set to the corresponding target signal sample. Since ideal sinc interpolation reproduces the data at its own interpolation nodes, the resulting reconstruction agrees with x at every point of the N^D lattice, yielding a zero optimal reconstruction error. The critical parameter budget is therefore $d^\star = N^D$ total parameters, corresponding to per-axis resolution $r^\star = N$. □

B.5 Proof of Theorem 4: FFN

Throughout this subsection, we consider an FFN [11] with Fourier feature embedding dimension q, a single-hidden layer ReLU MLP of width w, and input coordinates $v \in [0,1]^k$. The Fourier feature embedding is

$$\gamma_q(v) \;=\; \big[\; a_1 \cos(2\pi b_1^\top v),\; a_1 \sin(2\pi b_1^\top v),\; \ldots,\; a_q \cos(2\pi b_q^\top v),\; a_q \sin(2\pi b_q^\top v) \;\big]^\top \in \mathbb{R}^{2q}$$

where $b_l \sim \mathcal{N}(0, \sigma_b^2 I_k)$ and $a_l \sim \mathcal{N}(0, \sigma_a^2)$ for $l = 1, \ldots, q$. The FFN output is $g_d(\theta) = \mathrm{MLP}_\theta\big(\gamma_q(v)\big)$, where $d = P(L, w, q) := (L-1)w^2 + (2q + L + 1)w + 1$ denotes the total parameter count. In our analysis, we index the model family by the MLP width w while holding the embedding dimension q and number of hidden layers $L = 1$ fixed, unless stated otherwise. This is the regime in which the approximation-rate results from Siegel and Xu [42] and the lower-bound result of Achour et al. [41] can be applied most cleanly. For a target signal $x \in \mathbb{R}^n$ sampled at coordinates $v_1, \ldots, v_n$, the optimal reconstruction error at parameter budget d is

$$\|x - g^\star_d\| = \min_{g_d \in \mathcal{G}_d} \|x - g_d\| = \min_{\theta \in \mathbb{R}^d} \|x - g_d(\theta)\|,$$

where $\mathcal{G}_d = \{\, g_d(\theta) : \theta \in \mathbb{R}^d \}$ is the set of all outputs realizable by an FFN with parameter budget d.

Lemma F.1 (FFN satisfies Assumption A1)**.** *For all $d \geq 0$,*

$$\|x - g^\star_d\| \geq \|x - g^\star_{d+1}\|.$$

Proof. It suffices to show that $\mathcal{G}_d \subset \mathcal{G}_{d'}$ whenever $d' > d$. We consider three cases corresponding to the three ways of increasing the parameter count.

Case 1: Increasing the embedding dimension: $d' = P(L, w, q+1)$. An FFN with $q+1$ embedding frequencies can replicate any FFN with q embedding frequencies by appending the two new Fourier

features and setting the corresponding incoming weights in the first affine layer to zero. The realized function is unchanged by this operation, so $\mathcal{G}_d \subset \mathcal{G}_{d'}$.

Case 2: Increasing the hidden width: $d' = P(L, w+1, q)$. An FFN with hidden width $w+1$ can replicate any width-w FFN by padding the first-layer weight matrix with an all-zero row and extending subsequent weight matrices and bias vectors with zeros to maintain consistent dimensions. The additional neuron contributes zero to the output, and the realized function is identical. Thus, $\mathcal{G}_d \subset \mathcal{G}_{d'}$.

Case 3: Increasing the number of hidden layers: $d' = P(L+1, w, q)$. An FFN with $L+1$ hidden layers can replicate any L-layer FFN by inserting an identity layer between any existing ReLU and the next affine layer. The input to the inserted layer is the output of a preceding ReLU and is therefore nonnegative; setting the inserted layer's weights to the identity and its biases to zero, the subsequent ReLU acts as the identity on this nonnegative input. The realized function remains the same, so $\mathcal{G}_d \subset \mathcal{G}_{d'}$.

In all three cases, $\mathcal{G}_d \subset \mathcal{G}_{d'}$, so the infimum of the reconstruction error over the strictly larger set $\mathcal{G}_{d'}$ is no greater than over $\mathcal{G}_d$. Therefore, the optimal error is nonincreasing with parameter budget. □

Lemma F.2 (FFN admits a convex and monotone upper envelope for Assumption (A2))**.** *Fix the Fourier feature embedding dimension q and consider one-hidden-layer ReLU FFNs of width w. Let $\Omega \subset \mathbb{R}^{2q}$ be a bounded domain containing the range of the embedding γ_q, and assume there exists a function $f : \Omega \to \mathbb{R}$ in the Barron space $\mathcal{B}^1$ (i.e. $\int_{\mathbb{R}^{2q}} (1+|\omega|)|\hat{f}(\omega)|d\omega$ is bounded) such that $x_i = f(\gamma_q(v_i))$ for all $i = 1, \ldots, n$. Then the approximation error is bounded above by a monotone, discretely convex envelope of the form*

$$U(w) = Kw^{-\frac{1}{2}},$$

or equivalently,

$$U(d) = C_u d^{-\frac{1}{2}}.$$

Here $K > 0$ is a constant depending on the function f and the embedding range Ω, but independent of w. The constant $C_u > 0$ is the corresponding constant after rewriting the width-dependent bound in terms of the parameter count d.

Proof. Let $\mathcal{F}_w^{2q}(\sigma)$ be the class of width-w single-hidden-layer ReLU networks on $\mathbb{R}^{2q}$, where σ denotes the ReLU activation function. Following the norm convention stated at the beginning of the appendix, $\|f - f_w\|_{L^2(\Omega)}$ agrees with $\|x - g_w(\theta)\|_{\ell_2}$ up to a fixed sample-grid normalization. It follows that the optimal error at width w satisfies

$$\min_\theta \|x - g_w(\theta)\| \propto \inf_{f_w \in \mathcal{F}_w^{2q}(\sigma)} \|f - f_w\|_{L^2(\Omega)}.$$

Applying Corollary 1 (see below) with $n_0 = 3$ for the ReLU activation function [42], the L^2 approximation error satisfies

$$\inf_{f_w \in \mathcal{F}_w^{2q}(\sigma)} \|f - f_w\|_{L^2(\Omega)} = \inf_{f_w \in \mathcal{F}_w^{2q}(\sigma)} \|f - f_w\|_{H^0(\Omega)} \le |\Omega|^{\frac{1}{2}} C(c, 0, \mathrm{diam}(\Omega), \sigma)\sqrt{3}\|f\|_{\mathcal{B}^1} w^{-\frac{1}{2}},$$

for some $c > 0$. We can rewrite this as

$$\min_\theta \|x - g_w(\theta)\| \le U(w) = Kw^{-\frac{1}{2}},$$

for some $K > 0$ that depends on $\|f\|_{\mathcal{B}^1}$, $|\Omega|^{\frac{1}{2}}$, $C(c, 0, \mathrm{diam}(\Omega), \sigma)$, and the fixed normalization, but is independent of w. Note that $w^{-\frac{1}{2}}$ is convex on $(0, \infty)$. Thus, $U(w)$ is discretely convex in w. For $L = 1$ and fixed q, the parameter count $d = (2q+2)w + 1$ is affine in w. Thus, the upper bound can be rewritten as a function of d:

$$U(d) = C_u d^{-\frac{1}{2}},$$

and $U(d)$ inherits monotonicity and discrete convexity from $U(w)$. □

Theorem 5 (Siegel and Xu, 2020 [42])**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. If the activation function $\sigma \in W^{m,\infty}(\mathbb{R})$ is non-zero and satisfies the polynomial decay condition*

$$|\sigma^{(k)}(t)| \le C_p(1+|t|)^{-p} \tag{6}$$

for $0 \leq k \leq m$ *and some* $p > 1$*, we have*

$$\inf_{f_n \in \mathcal{F}_d^n(\sigma)} \|f - f_n\|_{H^m(\Omega)} \leq |\Omega|^{\frac{1}{2}} C(p, m, \operatorname{diam}(\Omega), \sigma) n^{-\frac{1}{2}} \|f\|_{\mathcal{B}^{m+1}},$$

for any $f \in \mathcal{B}^{m+1}$.

Corollary 1 (Siegel and Xu, 2020 [42])**.** *Let* $W_{loc}^{m,\infty}(\mathbb{R})$ *denote the space of functions whose weak derivatives up to order* m *are essentially bounded on every compact subset of* $\mathbb{R}$*. Let* $\sigma \in W_{loc}^{m,\infty}(\mathbb{R})$ *be an activation function and suppose that there exists a* $\nu \in \mathcal{F}_1^{n_0}(\sigma)$ *which satisfies the polynomial decay condition* (6) *in Theorem 5, we have*

$$\inf_{f_n \in \mathcal{F}_d^n(\sigma)} \|f - f_n\|_{H^m(\Omega)} \leq |\Omega|^{\frac{1}{2}} C(p, m, \operatorname{diam}(\Omega), \sigma) \sqrt{n_0} \|f\|_{\mathcal{B}^{m+1}} n^{-\frac{1}{2}}.$$

This follows immediately from Theorem 5 and the observation that $\nu \in \mathcal{F}_1^{n_0}(\sigma)$ implies that

$$\mathcal{F}_d^n(\nu) \subset \mathcal{F}_d^{nn_0}(\sigma).$$

Lemma F.3 (FFN admits a lower envelope for Assumption A2)**.** *Fix the Fourier feature embedding dimension* q *and consider one-hidden-layer ReLU FFNs of width* w*. Let* $1 \leq p < +\infty$ *and* $\mathcal{X}$ *be a measurable space endowed with a probability measure* μ*. Let* F *be a set of measurable functions from* $\mathcal{X}$ *to* $[-1, 1]$ *such that* $\log M(\varepsilon, F, \|\cdot\|_{L^p(\mu)}) \geq c_0 \varepsilon^{-\alpha}$ *for some* $c_0, \varepsilon_0, \alpha > 0$ *and all* $\varepsilon \leq \varepsilon_0$*. Here,* $M(\varepsilon, F, \|\cdot\|_{L^p(\mu)})$ *denotes the* ϵ*-packing number of* F *under* $\|\cdot\|_{L^p(\mu)}$*, defined as the maximum number of points in* F *that are pairwise more than* ϵ *apart. Then, there exist a positive integer* $d_{\min}$ *and* $c_1 > 0$ *such that, for any* $d \geq d_{\min}$*, the approximation error is bounded below by a monotonically decreasing envelope*

$$L(d) = c_1 d^{-\frac{1}{\alpha}} \log^{-\frac{3}{\alpha}}(d).$$

Proof. Define the FFN function class at width w as

$$\mathcal{G}_w = \{v \to \psi(\gamma_q(v)) : \psi \in \Psi_{w,2q}\},$$

where $\Psi_{w,2q}$ denotes the class of single-hidden-layer, width-w ReLU networks taking inputs in $\mathbb{R}^{2q}$. We first relate the complexity of $\mathcal{G}_w$ to the complexity of $\Psi_{w,2q}$. For a class S of real-valued functions on a domain Z, a set of points $\{z_1, \ldots, z_m\} \subseteq Z$ is said to be pseudo-shattered by S if there exists $r_1, \ldots, r_m \in \mathbb{R}$ such that for every sign pattern $o \in \{-1, +1\}^m$ there exists $s \in S$ satisfying $\operatorname{sign}(s(z_i) - r_i) = o_i$ for all $i = 1, \ldots, m$. The pseudo-dimension $\operatorname{Pdim}(S)$ is the largest integer m such that some m points are pseudo-shattered by S.

A related notion is the fat-shattering dimension [41]. For $\gamma > 0$, a set of points $\{z_1, \ldots, z_m\} \subseteq Z$ is γ-shattered by S if there exists $r_1, \ldots, r_m \in \mathbb{R}$ such that for every sign pattern $o \in \{-1, +1\}^m$ there exists $s \in S$ satisfying $s(z_i) \geq r_i + \gamma$ when $o_i = +1$ and $s(z_i) \leq r_i - \gamma$ when $o_i = -1$ for $i = 1, \ldots, m$. The γ-fat-shattering dimension $\operatorname{fat}_\gamma(S)$ is the largest integer m such that some m points are γ-shattered by S, with the convention $\operatorname{fat}_\gamma(S) = 0$ if no such set exists and $\operatorname{fat}_\gamma(S) = +\infty$ if arbitrarily large such sets exist. The fat-shattering dimension is dominated by the pseudo-dimension, $\operatorname{fat}_\gamma(S) \leq \operatorname{Pdim}(S)$ for all $\gamma > 0$. Thus, once we show that $\mathcal{G}_w$ has finite pseudo-dimension, the condition $\operatorname{Pdim}(G) < +\infty$ required by Theorem 6 follows automatically.

We claim

$$\operatorname{Pdim}(\mathcal{G}_w) \leq \operatorname{Pdim}(\Psi_{w,2q}).$$

Indeed, if a set of points $\{v_i\}_{i=1}^n \subset \mathcal{X}$ is pseudo-shattered by $\mathcal{G}_w$ with $(r_1, \ldots, r_n)$, then for every $o \in \{-1, +1\}^n$, there exists $g \in \mathcal{G}_w$ with $\operatorname{sign}(g(v_i) - r_i) = o_i$ for all i. By constructing $g = \psi \circ \gamma_q$ for some $\psi \in \Psi_{w,2q}$, the corresponding embedded points $\{\gamma_q(v_i)\}_{i=1}^n \subset \mathbb{R}^{2q}$ are then pseudo-shattered by $\Psi_{w,2q}$ with the same $(r_1, \ldots, r_n)$, so $n \leq \operatorname{Pdim}(\Psi_{w,2q})$. It follows that $\operatorname{Pdim}(\mathcal{G}_w) \leq \operatorname{Pdim}(\Psi_{w,2q})$. This argument does not require the Fourier embedding to preserve the full complexity of the decoder class; it only uses the fact that composing with a fixed embedding cannot increase pseudo-dimension.

The number of trainable parameters in $\Psi_{w,2q}$ is $d = (2q+2)w+1$. By the classical pseudo-dimension bound for ReLU networks with d weights and depth $L = 1$ [56], $\operatorname{Pdim}(\Psi_{w,2q}) = O(d \log d)$, or equivalently, $\operatorname{Pdim}(\Psi_{w,2q}) \leq Cd \log d$ for a universal constant C. Combining this with the previous inequality gives

$$\operatorname{Pdim}(\mathcal{G}_w) \leq Cd \log d.$$

Applying Theorem 6 to the function class $\mathcal{G}_w$ yields

$$\sup_{f\in F}\inf_{g\in\mathcal{G}_w}\|f-g\|_{L^p(\mu)}\geq\min\left\{\varepsilon_1,c_2\text{Pdim}(\mathcal{G}_w)^{-\frac{1}{\alpha}}\log^{-\frac{2}{\alpha}}(\text{Pdim}(\mathcal{G}_w))\right\},$$

for some constants $\epsilon_1, c_2 > 0$. Let $d_{\min}$ be a positive integer such that for any $d > d_{\min}$, the second term is smaller than ϵ_1. Since $t^{-\frac{1}{\alpha}}\log^{-\frac{2}{\alpha}}(t)$ is decreasing in t, we can substitute the bound $\text{Pdim}(\mathcal{G}_w)\leq Cd\log d$ to obtain

$$\sup_{f\in F}\inf_{g\in\mathcal{G}_w}\|f-g\|_{L^p(\mu)}\geq c_2(Cd\log d)^{-\frac{1}{\alpha}}\log^{-\frac{2}{\alpha}}(Cd\log d).$$

Absorbing all logarithmic constants c_1 yields

$$\sup_{f\in F}\inf_{g\in\mathcal{G}_w}\|f-g\|_{L^p(\mu)}\geq c_1d^{-\frac{1}{\alpha}}\log^{-\frac{3}{\alpha}}(d).$$

In other words, $L(d)=c_1d^{-\frac{1}{\alpha}}\log^{-\frac{3}{\alpha}}(d)$ is the desired lower envelope for the global representation error over the target class F. □

Theorem 6 (Achour, 2022 [41])**.** *Let $1\leq p<+\infty$ and $\mathcal{X}$ be a measurable space endowed with a probability measure μ. Let F, G be two sets of measurable functions from $\mathcal{X}$ to $\mathbb{R}$, such that all functions in F have the same $[a,b]$ for some $a<b$, and such that $fat_\gamma(G)<+\infty$ for all $\gamma>0$. If $\log M(\varepsilon,F,\|\cdot\|_{L^p(\mu)})\geq c_0\varepsilon^{-\alpha}$ for some $c_0,\varepsilon_0,\alpha>0$ and all $\varepsilon\leq\varepsilon_0$, and if Pdim$(G)<+\infty$, then there exist constants $c_1,\varepsilon_1>0$ depending only on $b-a,p,c_0,\varepsilon_0$ and α such that*

$$\sup_{f\in F}\inf_{g\in G}\|f-g\|_{L^p(\mu)}\geq\min\left\{\varepsilon_1,c_1\textit{Pdim}(G)^{-\frac{1}{\alpha}}\log^{-\frac{2}{\alpha}}(\textit{Pdim}(G))\right\}.$$

Here Pdim(G) denotes the pseudo-dimension of G.

Lemma F.4 (FFN satisfies Assumption A3)**.** *There exists a finite parameter budget $d^\star$ such that $\|x-g^\star_{d^\star}\|=0$ with high probability.*

Proof. The proof proceeds in two steps. First, we show that there exists a finite embedding dimension $q^\star$ such that the Fourier feature matrix has full row rank, guaranteeing exact interpolation of the target signal at all sample points. Second, we invoke the minimum-width universal approximation results to show that a single-hidden-layer ReLU MLP of finite width can realize the required linear map on the embedded coordinates.

Step 1: Full row rank of the Fourier feature matrix. Let $S=\{v_1,v_2,\dots,v_n\}$ be the finite set of distinct sampled coordinates at which the signal x is defined, and form the Fourier feature matrix

$$\Phi_S:=\begin{bmatrix}\gamma_q(v_1)^\top\\ \vdots\\ \gamma_q(v_n)^\top\end{bmatrix}\in\mathbb{R}^{n\times 2q}.$$

We claim that there exists $q^\star$ with $2q^\star\geq n$ such that $\text{rank}(\Phi_S)=n$ with high probability over the random draws of the embedding frequencies. To establish this, factor $\Phi_S=\Phi'_S\,\text{diag}(a_1,a_1,\dots,a_q,a_q)$ where

$$\Phi'_S=\begin{bmatrix}\cos(2\pi b_1^\top v_1),\ \sin(2\pi b_1^\top v_1) & \cdots & \cos(2\pi b_q^\top v_1),\ \sin(2\pi b_q^\top v_1)\\ \vdots & \ddots & \vdots\\ \cos(2\pi b_1^\top v_n),\ \sin(2\pi b_1^\top v_n) & \cdots & \cos(2\pi b_q^\top v_n),\ \sin(2\pi b_q^\top v_n)\end{bmatrix}.$$

Since $a_l\sim\mathcal{N}(0,\sigma_a^2)$ for $l=1,\dots,q$, all $2q$ entries on the diagonal are nonzero almost surely, so the diagonal matrix $\text{diag}(a_1,a_1,\dots,a_q,a_q)$ has full rank with probability one. Hence, $\text{rank}(\Phi_S)=\text{rank}(\Phi'_S)$ with probability one and it suffices to show that Φ'_S has full row rank with high probability. Consider the Gram matrix $G=\Phi'_S(\Phi'_S)^\top=\sum_{l=1}^q G_l$, where each $G_l\in\mathbb{R}^{n\times n}$ is the symmetric positive semidefinite matrix with entries $G_l[i,j]=\cos(2\pi b_l^\top(v_i-v_j))$. We note that $G_l=U_lU_l^\top$ where

$$U_l=\begin{bmatrix}\cos(2\pi b_l^\top v_1) & \sin(2\pi b_l^\top v_1)\\ \cos(2\pi b_l^\top v_2) & \sin(2\pi b_l^\top v_2)\\ \vdots & \vdots\\ \cos(2\pi b_l^\top v_n) & \sin(2\pi b_l^\top v_n)\end{bmatrix}\in\mathbb{R}^{n\times 2}.$$

It follows that $G_l \succeq 0$ and $\|G_l\| \leq \|U_l\|^2 \leq \|U_l\|_F^2 = tr(U_l^\top U_l) = n$. Let $\Sigma = \mathbb{E}[G_l]$. Then the (i,j) entry of Σ is

$$\Sigma[i,j] = \mathbb{E}[\cos(2\pi b_l^\top (v_i - v_j))] = \exp\left(-2\pi^2\sigma_b^2\|v_i - v_j\|_2^2\right),$$

which is the Gaussian RBF kernel evaluated at $v_i - v_j$. Since the points $v_1, \ldots, v_n$ are distinct, Σ is a Gram matrix of a Gaussian RBF kernel on distinct points and is therefore strictly positive definite. Let $\lambda = \lambda_{\min}(\Sigma) > 0$ be the smallest eigenvalue of Σ.

Define the centered, rescaled matrices $W_l = \Sigma^{-1/2} G_l \Sigma^{-1/2}$. Then W_l is symmetric and positive semidefinite by construction. We also have that

$$\mathbb{E}[W_l] = \Sigma^{-1/2}\mathbb{E}[G_l]\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_n, \qquad \text{and}$$

$$\|W_l\| \leq \|\Sigma^{-1/2}\|\|G_l\|\|\Sigma^{-1/2}\| = \|\Sigma^{-1/2}\|^2\|G_l\| = \left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}\right)^2 \|G_l\| \leq \frac{n}{\lambda}.$$

Let the diagonalization of W_l be $W_l = PDP^\top$, where P is orthogonal and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. By definition, for all $i \in [n]$, $\lambda_i \leq \lambda_{\max}(W_l) = \|W_l\|$ and $\lambda_i^2 \leq \|W_l\|\lambda_i$. It follows that $D^2 \preceq \|W_l\| D$ and $W_l^2 = PD^2P^\top \preceq \|W_l\| PDP^\top = \|W_l\| W_l$. Define $G_l' = \frac{1}{q}(W_l - I)$. We have

$$\|G_l'\| \leq \frac{1}{q}(\|W_l\| + \|I_n\|) \leq \frac{1}{q}\left(\frac{n}{\lambda} + 1\right),$$

$$\mathbb{E}[G_l'] = \frac{1}{q}(\mathbb{E}[W_l] - I_n) = 0_n, \qquad \text{and}$$

$$\begin{aligned}
\mathbb{E}[(G_l')^2] &= \frac{1}{q^2}\mathbb{E}[W_l^2 - 2W_l + I_n] \\
&= \frac{1}{q^2}\mathbb{E}[W_l^2] - 2I_n + I_n \\
&\preceq \frac{1}{q^2}\left(\|W_l\|\mathbb{E}[W_l] - I_n\right) \\
&\preceq \frac{1}{q^2}\left(\frac{n}{\lambda} - 1\right) I_n \preceq \frac{n}{q^2\lambda} I_n.
\end{aligned}$$

Thus, G_l' are independent centered self-adjoint random $n \times n$ matrices with

$$\|G_l'\| \leq \frac{1}{q}\left(\frac{n}{\lambda} + 1\right) \qquad \text{and} \qquad \left\|\sum_{l=1}^q \mathbb{E}[(G_l')^2]\right\| \leq \frac{n}{q\lambda}.$$

Applying the non-commutative Bernstein inequality (Theorem 7) to the G_l' matrices, we have that for every $\epsilon \geq 0$,

$$\delta = \mathbb{P}\left\{\|\sum_{l=1}^q G_l'\| \geq \epsilon\right\} \leq 2n \cdot \exp\left(\frac{-\epsilon^2/2}{\frac{n}{q\lambda} + \frac{(n/\lambda+1)\epsilon}{3q}}\right) = 2n \cdot \exp\left(\frac{-\epsilon^2 q}{\frac{n}{\lambda}\cdot\frac{2\epsilon+6}{3} + \frac{2\epsilon}{3}}\right).$$

Pick $\epsilon = 1/2$. We get

$$\mathbb{P}\left\{\|\sum_{l=1}^q G_l'\| < \frac{1}{2}\right\} = 1 - \delta \geq 1 - 2n \cdot \exp\left(\frac{-q}{\frac{n}{\lambda}\cdot\frac{28}{3} + \frac{4}{3}}\right).$$

This implies that if $q \geq C\frac{n}{\lambda}\ln\frac{2n}{\delta}$ for some absolute constant C, then $\mathbb{P}\left\{\|\sum_{l=1}^q G_l'\| < \frac{1}{2}\right\} \geq 1 - \delta$. We can write

$$\|\sum_{l=1}^q G_l'\| < \frac{1}{2} \iff \|\sum_{l=1}^q \left(\frac{1}{q}W_l - \frac{1}{q}I_n\right)\| < \frac{1}{2} \iff \|\left(\frac{1}{q}\sum_{l=1}^q W_l\right) - I_n\| < \frac{1}{2}.$$

Since W_l is symmetric, $\left(\frac{1}{q}\sum_{l=1}^q W_l\right) - I_n$ is also symmetric and $\left|\lambda_{\max}\left(\left(\frac{1}{q}\sum_{l=1}^q W_l\right) - I_n\right)\right| = \|\left(\frac{1}{q}\sum_{l=1}^q W_l\right) - I_n\|$. It follows that $\left|\lambda_{\max}\left(\left(\frac{1}{q}\sum_{l=1}^q W_l\right) - I_n\right)\right| < \frac{1}{2}$ and all eigenvalues of

$\left(\frac{1}{q}\sum_{l=1}^{q} W_l\right) - I_n$ lie in $(-\frac{1}{2}, \frac{1}{2})$. As a result, all eigenvalues $\frac{1}{q}\sum_{l=1}^{q} W_l$ lies in $(\frac{1}{2}, \frac{3}{2})$. It follows that

$$\begin{aligned}\frac{1}{q}\sum_{l=1}^{q} W_l \succeq \frac{1}{2} I_n &\iff \Sigma^{-1/2}\left(\sum_{l=1}^{q} G_l\right)\Sigma^{-1/2} \succeq \frac{q}{2} I_n \\ &\iff G \succeq \frac{q}{2}\Sigma \\ &\iff \lambda_{\min}(G) \geq \frac{q}{2}\lambda > 0.\end{aligned}$$

Therefore, we can conclude that for $q \geq C\frac{n}{\lambda}\ln\frac{2n}{\delta}$ for some absolute constant C, then $\operatorname{rank}(\Phi_S') = \operatorname{rank}(G) = n$ with probability at least $1-\delta$.

Since Φ_S has full row rank, there exists a vector $c \in \mathbb{R}^{2q^\star}$ such that $\Phi_S c = x_S$ (for instance, the minimum-norm solution $c = \Phi_S^\top(\Phi_S\Phi_S^\top)^{-1}x_S$)). The linear map $v \to c^\top\gamma_{q^\star}(v)$ therefore interpolates x exactly at all n sampled points.

Step 2: Realization by a finite-width ReLU MLP. It remains to show that the linear map $z \to c^\top z$ can be realized by a single-hidden-layer ReLU MLP of finite width. This is immediate: a width-$4q^\star$ network with weights

$$W_1 = \begin{bmatrix} I_{2q^\star} \\ -I_{2q^\star}\end{bmatrix} \in \mathbb{R}^{4q^\star\times 2q^\star}, \quad b_1 = 0 \in \mathbb{R}^{4q^\star}, \quad W_{out} = \begin{bmatrix} c^\top & -c^\top\end{bmatrix} \in \mathbb{R}^{1\times 4q^\star}, \quad b_{out} = 0 \in \mathbb{R}$$

satisfies $W_{out}\cdot\operatorname{ReLU}(W_1 z) = c^\top(\operatorname{ReLU}(z) - \operatorname{ReLU}(-z)) = c^\top z$ for all z, since $\operatorname{ReLU}(z) - \operatorname{ReLU}(-z) = z$. In particular, evaluating at $z = \gamma_{q^\star}(v_i)$ yields $c^\top\gamma_{q^\star}(v_i) = x_{S,i}$ for all $i = 1,\ldots,n$.

More generally, for the effective input dimension $d_x = 2q$ and output dimension $d_y = 1$,the minimum width for L^p-approximation of any function in $L^p([0,1]^{d_x}, \mathbb{R}^{d_y})$ by ReLU networks with at least 2 hidden layers is $w_{\min} = \max\{d_x, d_y, 2\}$, as established by Kim et al. [47]. This guarantees that any continuous function on the compact domain of the Fourier embedding, including the above interpolant, can be approximated to arbitrary precision by a ReLU MLP of at least 2 hidden layers and width at least $2q$. The critical parameter budget is therefore either $d^\star = P(L=1, w=4q^\star, q=q^\star)$ or $d^\star = P(L=2, w=2q^\star, q=q^\star)$, which are both finite. □

Theorem 7 (Non-commutative Bernstein-type inequality (Vershynin, 2011 [57]))**.** *Consider a finite sequence X_i of independent centered self-adjoint random $n\times n$ matrices. Assume we have for some numbers K and σ that*

$$\|X_i\| \leq K \quad \textit{almost surely,} \qquad \Big\|\sum_i \mathbb{E}X_i^2\Big\| \leq \sigma^2.$$

Then, for every $t \geq 0$ we have

$$\mathbb{P}\left\{\Big\|\sum_i X_i\Big\| \geq t\right\} \leq 2n\cdot\exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3}\right).$$

B.6 Proof of Theorem 4: Instant-NGP

We consider an Instant-NGP [13] model with L resolution levels, T entries per level hash table, F feature dimensions per entry, and a single-hidden-layer ReLU MLP decoder of width w. The number of trainable encoding parameters is TLF, and the total parameter count $d = TLF + d_{MLP}$ includes both encoding and decoder parameters. Define the full concatenated encoding at coordinate $v \in \mathbb{R}^k$ as

$$z(v) = [e_1(v) \quad e_2(v) \quad \cdots \quad e_L(v) \quad \xi(v)] \in \mathbb{R}^{LF+E},$$

where $e_l(v)$ is the interpolated hash-grid feature at level l and $\xi(v) \in \mathbb{R}^E$ denotes an auxiliary non-spatial encoding. Following [13], we index the model family by the encoding hyperparameters (T, L) while holding F, the spatial input dimension k, the coarsest resolution $N_{\min}$, and the decoder architecture fixed. For a target signal $x \in \mathbb{R}^n$ sampled at coordinates $v_1,\ldots,v_n \in \mathbb{R}^k$, the optimal reconstruction error at parameter budget d is

$$\|x - g_d^\star\| = \min_{\theta,\Phi}\|x - g_d(\theta,\Phi)\|,$$

where θ denotes the hash table feature parameters and Φ encompasses the MLP weights and biases.

To analyze the error curve, we adopt two standard structural assumptions on the spatial hash function used in the multi-resolution encoding [58]:

- **Uniform Hashing.** Let $\hat{h}_T$ denote the spatial hash function used by the multi-resolution encoding. Then $\Pr[\hat{h}_T(u) = \hat{h}_T(v)] \leq \frac{1}{T}$ for any two distinct integer vertices $u \neq v$.
- **Independence Across Levels.** Hash collision events at different resolution levels are mutually independent.

These assumptions reflect the design objectives of the per-dimension linear-congruential XOR hash used in Instant-NGP, namely decorrelating the contribution of the input dimensions and approximately uniform distribution over the table.

Lemma I.1 (Instant-NGP satisfies Assumption A1)**.** *For all parameter budgets $d \geq 0$,*

$$\|x - g_d^\star\| \geq \|x - g_{d+1}^\star\|.$$

Proof. It suffices to show that the model class $\mathcal{G}_d$ is contained in $\mathcal{G}_{d'}$ whenever $d' > d$. We first consider increasing the number of levels. A model with L levels computes an encoding of the form

$$z(v) = [e_1(v) \quad e_2(v) \quad \cdots \quad e_L(v) \quad \xi(v)] \in \mathbb{R}^{LF+E},$$

where $e_l(v)$ is the interpolated hash-grid feature at level l and $\xi(v) \in \mathbb{R}^E$ denotes any auxiliary non-spatial encoding. Let the first affine layer of the decoder have weights

$$W = [W_{\text{hash}} \quad W_\xi],$$

where W_{hash} acts on the LF-dimensional concatenated hash features. A model with $L+1$ levels has encoding

$$z(v) = [e_1(v) \quad e_2(v) \quad \cdots \quad e_L(v) \quad e_{L+1}(v) \quad \xi(v)] \in \mathbb{R}^{(L+1)F+E}.$$

It can reproduce the original L-level model by choosing the new first-layer weights as

$$W = [W_{\text{hash}} \quad 0 \quad W_\xi],$$

setting the new incoming weights from level $L+1$ to zero and keeping all remaining decoder weights unchanged. The additional level may contain arbitrary feature values, but it contributes nothing to the output. Thus every reconstruction realizable with L levels is also realizable with $L+1$ levels.

We now consider increasing the hash table size by an integer multiple. Let $T' = mT$ for an integer $m \geq 1$. At a fixed level, the spatial hash function has the form

$$\hat{h}_T(p) = H(p) \mod T,$$

where p is an integer grid vertex and $H(p) = \bigoplus_{i=1}^k p_i \pi_i$ is the pre-modulo hash value with π_i being unique, large prime numbers. Given any table $\theta^{(l)} \in \mathbb{R}^{T \cdot F}$ for the smaller model, define the larger table features $\theta'^{(l)} \in \mathbb{R}^{T' \cdot F}$ by

$$\theta_j'^{(l)} = \theta_{j \bmod T}^{(l)}$$

Since $T' = mT$, we have $\hat{h}_{T'}(p) \mod T = \hat{h}_T(p)$. Then, for any level l and corner p, the looked-up feature in the larger model $\theta'^{(l)}_{\hat{h}_{T'}(p)} = \theta^{(l)}_{\hat{h}_{T'}(p) \bmod T} = \theta^{(l)}_{\hat{h}_T(p)}$ matches the corresponding lookup in the smaller model. With this construction of the larger model, the interpolation weights depend only on the input coordinate and grid resolution, not on T, so the interpolated feature at every level is identical in the two models. Keeping the weights fixed gives the same output for every input coordinate v_i for $i = 1, \ldots, n$. Thus, $\mathcal{G}_T \subseteq \mathcal{G}_{mT}$.

In both cases (increasing the number of levels L or hash table size T), the model class is nested, so the infimum of reconstruction error over the larger class is no greater than over the smaller class. Therefore, $\|x - g_d^\star\|$ is nonincreasing in d. □

Lemma I.2 (For Assumption A2, Instant-NGP admits a monotone, discretely convex upper envelope for the expected optimal squared error, where the expectation is taken over the random hash function)**.**

Under the uniform hashing and level-independence assumptions, the expected squared optimal reconstruction error of Instant-NGP, taken over the random spatial hash function, satisfies

$$\mathbb{E}[\|x - g_d^\star\|^2] \le C_n \left(\frac{C_k}{T}\right)^L = U(d)^2,$$

where $C_n = \frac{1}{2}\binom{n}{2}, C_k = 4^k$, and $d^\star$ is the critical parameter budget identified in Lemma I.4 (below). The upper envelope $U(d)$ is monotone decreasing and discretely convex in T (with L, F fixed) and in L (with T, F fixed).

Proof. Let $U_l(v_i)$ be the set of 2^k integer grid vertices involved in the k-linear interpolation of v_i at level l. For the level-l interpolated feature vectors for v_i and v_j to be indistinguishable, at least one of the queried vertices for v_i must collide with one of the queried vertices for v_j. By union bound and the uniform hashing assumption,

$$\begin{aligned}
\Pr[A_{ij}^{(l)}] &\le \Pr\left[\bigcup_{u_i \in U_l(v_i)} \bigcup_{u_j \in U_l(v_j)} \{h(u_i) = h(u_j)\}\right] \\
&\le \sum_{u_i \in U_l(v_i)} \sum_{u_j \in U_l(v_j)} \Pr[h(u_i) = h(u_j)] \\
&\le |U_l(v_i)| \cdot |U_l(v_j)| \cdot \frac{1}{T} \\
&\le 2^k \cdot 2^k \cdot \frac{1}{T} = \frac{C_k}{T},
\end{aligned}$$

where $A_{ij}^{(l)}$ is the event that the level-l encoded features for v_i and v_j collide, and $C_k = 4^k$. Since the full encoding concatenates all L levels, the two samples are indistinguishable to the decoder only if they are indistinguishable at every level. By independence across levels,

$$\Pr[A_{ij}] = \prod_{l=1}^{L} \Pr[A_{ij}^{(l)}] \le \left(\frac{C_k}{T}\right)^L,$$

where $A_{ij} = \bigcap_{l=1}^{L} A_{ij}^{(l)}$ is the event that the interpolated feature vectors v_i and v_j are indistinguishable at all levels. Whenever A_{ij} occurs, the decoder receives the same input for v_i and v_j, and therefore every Instant-NGP reconstruction must assign the same predicted value to these samples. More generally, the relation of structural indistinguishability partitions $\{1, \dots, n\}$ into equivalence classes $C \in \Pi_{T,L}$, and every reconstruction is constant on each such class. The optimal MLP outputs the within-class mean $\bar{x}_C$ on each class C, so

$$\begin{aligned}
\|x - g_{T,L}^\star\|^2 &= \sum_{C \in \Pi_{T,L}} \sum_{i \in C} (x_i - \bar{x}_C)^2 \\
&= \sum_{C \in \Pi_{T,L}} \sum_{i \in C} (x_i - \frac{1}{|C|} \sum_{j \in C} x_j)^2 \\
&= \sum_{C \in \Pi_{T,L}} \frac{1}{|C|} \sum_{i,j \in C} (x_i - x_j)^2 \\
&= \sum_{C \in \Pi_{T,L}} \frac{1}{2|C|} \sum_{\substack{i,j \in C \\ i<j}} (x_i - x_j)^2.
\end{aligned}$$

Using $1/|C| \le 1 \ \forall C \in \Pi_{T,L}$ and the normalization $x_i \in [0,1]$, we can write

$$\|x - g_{T,L}^\star\|^2 \le \frac{1}{2} \sum_{C \in \Pi_{T,L}} \sum_{\substack{i,j \in C \\ i<j}} (x_i - x_j)^2 \le \frac{1}{2} \sum_{i<j} \mathbb{1}[A_{ij}].$$

Taking expectation over the random hash function gives the upper envelope:

$$\begin{aligned}\mathbb{E}[\|x - g^\star_{T,L}\|^2] &\le \frac{1}{2}\sum_{i<j}\Pr[A_{ij}] \\ &\le \frac{1}{2}\binom{n}{2}\left(\frac{C_k}{T}\right)^L, \quad \text{since there are } \binom{n}{2} \text{ pairs of input coordinates} \\ &= C_n\left(\frac{C_k}{T}\right)^L = U(T,L)^2, \quad \text{where } C_n \text{ is a constant depending on } n.\end{aligned}$$

It remains to verify the monotone discrete convexity of the upper envelope. For fixed L,

$$U(T,L) = CT^{-L/2}$$

for a constant $C > 0$. The function $f(t) = t^{-L/2}$ is decreasing and convex on $(0,\infty)$, for $L \ge 1$, since its first derivative is negative and its second derivative is positive. Therefore, its restriction to integer table sizes is discretely convex:

$$U(T-1,L) - U(T,L) \ge U(T,L) - U(T+1,L).$$

For fixed $T > C_k$, let $q = (C_k/T)^{1/2} \in (0,1)$. Then

$$U(T,L) = Cq^L$$

for a constant $C > 0$. Hence

$$U(T,L-1) - U(T,L) - (U(T,L) - U(T,L+1)) = C(q^{L-1} - 2q^L + q^{L+1}) = Cq^{L-1}(q-1)^2 > 0.$$

Thus the upper envelope is monotone decreasing and discretely convex in either T or L. □

Lemma I.3 (For Assumption A2, Instant-NGP admits a lower envelope for the expected optimal squared error, where the expectation is taken over the random hash function)**.** *Under the uniform hashing and level-independence assumptions, the expected squared optimal reconstruction error of Instant-NGP, taken over the random spatial hash function, satisfies*

$$\mathbb{E}[\|x - g^\star_d\|^2] \ge \delta_x \cdot \mathbb{1}\{d \le d^\star\} \cdot T^{-2^k L} = L(d)^2,$$

where $\delta_x > 0$ is a positive signal-dependent constant and $d^\star$ is the critical parameter budget identified in Lemma I.4 (below). The lower envelope $L(d)$ need not be convex.

Proof. In contrast to the upper-bound argument in Lemma I.2, which uses the necessary condition that indistinguishability at level l requires at least one collision among the queried corner vertices, the lower bound needs the stricter sufficient event that all corresponding corner vertices collide, which guarantees that the interpolated features are identical. This restriction is what allows the upper envelope and the lower bound to be expressed in compatible terms. Let $A^{(l)}_{ij}$ be the event that the level-l encoded features for v_i and v_j collide. Consider any pair (v_i, v_j) that lie at the same relative position within distinct voxels at every level l, so that the k-linear interpolation weights match. For such a pair, the event $A^{(l)}_{ij}$ reduces to the requirement that each of the 2^k corners of v_i's voxel hashes to the same bucket as the corresponding corner of v_j's voxel. Treating the 2^k corner pairs as independent uniform draws,

$$\Pr[A^{(l)}_{ij}] \ge \left(\frac{1}{T}\right)^{2^k},$$

and by independence across levels,

$$\Pr[A_{ij}] \ge \left(\frac{1}{T}\right)^{2^k L},$$

where $A_{ij} = \bigcap_{l=1}^{L} A^{(l)}_{ij}$ is the event that the interpolated feature vectors v_i and v_j are indistinguishable at all levels. Whenever A_{ij} occurs, the decoder receives the same input for v_i and v_j, and therefore every Instant-NGP reconstruction must assign the same predicted value to these samples. More

generally, the relation of structural indistinguishability partitions $\{1, \dots, n\}$ into equivalence classes $C \in \Pi_{T,L}$, and every reconstruction is constant on each such class. The optimal MLP outputs the within-class mean $\bar{x}_C$ on each class C, so

$$\begin{aligned}
\|x - g^\star_{T,L}\|^2 &= \sum_{C \in \Pi_{T,L}} \sum_{i \in C} (x_i - \bar{x}_C)^2 \\
&= \sum_{C \in \Pi_{T,L}} \sum_{i \in C} (x_i - \frac{1}{|C|} \sum_{j \in C} x_j)^2 \\
&= \sum_{C \in \Pi_{T,L}} \frac{1}{|C|} \sum_{i,j \in C} (x_i - x_j)^2 \\
&= \sum_{C \in \Pi_{T,L}} \frac{1}{2|C|} \sum_{\substack{i,j \in C \\ i<j}} (x_i - x_j)^2.
\end{aligned}$$

Using $1/|C| \geq 1/n \; \forall C \in \Pi_{T,L}$, we can write

$$\|x - g^\star_{T,L}\|^2 \geq \frac{1}{2n} \sum_{i<j} \mathbb{1}[A_{ij}](x_i - x_j)^2.$$

Let $\mathcal{P}$ be the set of pairs (i, j) at matching relative positions within their voxels and let $\delta_{\min} = \min_{(i,j) \in \mathcal{P}} (x_i - x_j)^2 > 0$ for a target signal whose sampled values are pairwise distinct on $\mathcal{P}$. Taking expectation over the random hash function yields

$$\mathbb{E}[\|x - g^\star_{T,L}\|^2] \geq \frac{\delta_{\min} |\mathcal{P}|}{2n} \cdot T^{-2^k L} = \delta_x \cdot T^{-2^k L},$$

where δ_x absorbs all signal-dependent constants. When $d \geq d^\star$, Lemma I.4 gives $\|x - g^\star_{T^\star, L^\star}\|^2 = 0$, so we set $L(d) = 0$ in this regime. Combining the two regimes yields a valid lower envelope

$$L(d) = \left(\delta_x \cdot \mathbb{1}\{d \leq d^\star\} \cdot T^{-2^k L} \right)^{1/2}.$$

□

Lemma I.4 (Instant-NGP satisfies Assumption A3)**.** *There exists a finite parameter budget $d^\star$ such that $\|x - g^\star_{d^\star}\| = 0$.*

Proof. Since the target signal is observed on finitely many sample coordinates $v_1, \dots, v_n$, choose the number of levels $L^\star$ such that finest grid resolution $N_{\max}$ is large enough for each sample coordinate to be represented by a distinct finest-level grid vertex. For grid-sampled images, volume, or radiance-field data, it suffices to take $N_{\max} \geq R$, where R denotes the resolution of x along each spatial axis.

Next, choose the finest-level hash table size so that $T^\star \geq N^k_{\max}$. Then every finest-level grid vertex can be assigned a unique table entry, so the finest level behaves as a dense feature grid with no hash collisions. That is, each input coordinate v_i is assigned a unique feature vector $\theta^{(L)}_i$ at level L, and the encoding is injective: $z(v_i) \neq z(v_j)$ for all $i \neq j$.

We now construct an exact representation. Set all decoder weights connected to coarser levels to zero and let the decoder depend only on the finest-level feature block. Assign the finest-level feature vector at the vertex corresponding to v_i so that a linear readout returns x_i. For scalar-valued signals, one feature channel is sufficient; concretely, set

$$\text{MLP}_\Phi(z(v_i)) = \text{MLP}_\Phi\left(\begin{bmatrix} \theta^{(0)}_i & \cdots & \theta^{(l)}_i & \cdots & \theta^{(L)}_i \end{bmatrix} \right) = \text{MLP}_\Phi\left(\begin{bmatrix} 0 & \cdots & 0 & \theta^{(L)}_i \end{bmatrix} \right) = x_i.$$

For vector-valued signals, one may either use one feature channel per output coordinate or a finite-width linear output layer. Since the Instant-NGP decoder ends in a linear layer, this readout can be implemented by the MLP by choosing the hidden layers to act as an identity map on the relevant feature coordinates, or equivalently by using the final affine layer to read out the stored value directly. With this construction, $g_{d^\star}(v_i) = x_i$ for every $i = 1, \dots, n$. Hence, the optimal error at the corresponding finite budget $d^\star = T^\star L^\star F + d_{MLP}$ is zero. □

B.7 Proof of Theorem 4: GA-Planes

We prove the GA-Planes [19] case for the two-dimensional nonconvex variant introduced in [19] and index the model family by its line feature dimension k_1, holding the line resolution r_1, plane resolution r_2, and plane feature dimension $k_2 = 1$ fixed. Let the target signal $x \in \mathbb{R}^n$ be arranged as a matrix $M \in \mathbb{R}^{N\times N}$ with $n = N^2$, and write the model output at integer coordinate (i, j) as

$$g_d(\theta)[i, j] = W_{\text{line}}^\top(e_1[i] \circ e_2[j]) + W_{\text{plane}}^\top e_{12}[i, j, :] + b,$$

where e_1, e_2 are linearly interpolated from line feature grids $g_1, g_2 \in \mathbb{R}^{r_1\times k_1}$, e_{12} is bilinearly interpolated from a plane feature grid $g_{12} \in \mathbb{R}^{r_2\times r_2\times k_2}$, $\circ$ denotes elementwise multiplication along the feature dimension, and $\theta = (g_1, g_2, g_{12}, W_{\text{line}}, W_{\text{plane}}, b)$ are the trainable parameters.

A central tool in the analysis is the matrix-completion equivalence established in [19] for the 2D linear-decoder setting: the reconstructed matrix $\hat{M} \in \mathbb{R}^{N\times N}$ produced by g_d admits the decomposition

$$\hat{M} = UV^T + \varphi(L),$$

with $U = g_1 \operatorname{diag}(W_{\text{line}}) \in \mathbb{R}^{N\times k_1}, V = g_2 \in \mathbb{R}^{N\times k_1}, L \in \mathbb{R}r_2r_2$ encoding the rescaled and bias-shifted plane grid, and $\varphi : \mathbb{R}^{r_2\times r_2} \to \mathbb{R}^{N\times N}$ the bilinear upsampling operator. The optimal squared compression loss at line feature dimension k_1 therefore equals

$$\|x - g_{k_1}^\star\|^2 = \min_{U,V,L} \|M - (UV^T + \varphi(L))\|_F^2. \tag{7}$$

This casts the analysis of 2D GA-Planes as a low-rank plus low-resolution matrix completion problem.

Lemma GAP.1 (GA-Planes satisfies Assumption A1)**.** *For all $k_1 \geq 0$,*

$$\|x - g_{k_1}^\star\| \geq \|x - g_{k_1+1}^\star\|.$$

Proof. Let $\mathcal{G}_{k_1}$ denote the set of model outputs realizable at line feature dimension k_1. It suffices to show $\mathcal{G}_{k_1} \subseteq \mathcal{G}_{k_1+1}$. Let $g \in \mathcal{G}_{k_1}$ be represented by the parameters $(g_1, g_2, g_{12}, W_{\text{line}}, W_{\text{plane}}, b)$. Construct a model in $\mathcal{G}_{k_1+1}$ by zero-padding the line components

$$g_1' = [g_1, 0] \in \mathbb{R}^{r_1\times(k_1+1)}, \quad g_2' = [g_2, 0] \in \mathbb{R}^{r_1\times(k_1+1)}, \quad W_{\text{line}}' = \begin{bmatrix} W_{\text{line}} \\ 0 \end{bmatrix} \in \mathbb{R}^{k_1+1}$$

while keeping the plane and bias components unchanged: $g_{12}' = g_{12}, W_{\text{plane}}' = W_{\text{plane}}, b' = b$. Since linear interpolation is applied channelwise, the new interpolated line features satisfy

$$e_1' = \begin{bmatrix} e_1 \\ 0 \end{bmatrix}, \quad e_2' = \begin{bmatrix} e_2 \\ 0 \end{bmatrix}, \quad \text{and hence} \quad (e_1' \circ e_2') = \begin{bmatrix} (e_1 \circ e_2) \\ 0 \end{bmatrix}.$$

The appended channel of W_{line}' is zero, so

$$W_{\text{line}}'^\top(e_1' \circ e_2') + W_{\text{plane}}'^\top e_{12}' + b' = W_{\text{line}}^\top(e_1 \circ e_2) + W_{\text{plane}}^\top e_{12} + b,$$

which equals the output of $g \in \mathcal{G}_{k_1}$. Therefore, $\mathcal{G}_{k_1} \subseteq \mathcal{G}_{k_1+1}$ and the optimal reconstruction error is nonincreasing in k_1. □

Lemma GAP.2 (GA-Planes satisfies Assumption A2 for squared error)**.** *For all $k_1 \geq 0$,*

$$\|x - g_{k_1}^\star\|^2 - \|x - g_{k_1+1}^\star\|^2 \geq \|x - g_{k_1+1}^\star\|^2 - \|x - g_{k_1+2}^\star\|^2.$$

Proof. By the matrix-completion equivalence in eq. (7), the squared optimal error decomposes as a joint minimization over a rank-$\leq k_1$ component $UV^\top$ and a low-resolution component $\varphi(L)$. Following [19], for fixed r_2, this joint optimization is equivalent to fitting the optimal low-resolution component first and then approximating the residual with a rank-$\leq k_1$ matrix. Let $L^\star$ be the optimal low-resolution fit and define the residual

$$R = M - \varphi(L^\star).$$

Substituting $UV^\top$ as the rank-k_1 approximation of R into eq. (7) and applying the Eckart-Young theorem yields

$$\|x - g_{k_1}^\star\|^2 = \min_{\operatorname{rank}(A)\leq k_1} \|R - A\|_F^2 = \sum_{i>k_1} \sigma_i(R)^2,$$

where $\sigma_1(R) \geq \sigma_2(R) \geq \cdots \geq 0$ are the singular values of R. We can write the consecutive differences as

$$\|x - g^\star_{k_1}\|^2 - \|x - g^\star_{k_1+1}\|^2 = \sum_{i>k_1} \sigma_i(R)^2 - \sum_{i>k_1+1} \sigma_i(R)^2 = \sigma_{k_1+1}(R)^2$$

and

$$\|x - g^\star_{k_1+1}\|^2 - \|x - g^\star_{k_1+2}\|^2 = \sum_{i>k_1+1} \sigma_i(R)^2 - \sum_{i>k_1+2} \sigma_i(R)^2 = \sigma_{k_1+2}(R)^2.$$

Since the singular values are nonincreasing, $\sigma_{k_1+1}(R) \geq \sigma_{k_1+2}(R)$, and so

$$(\|x - g^\star_{k_1}\|^2 - \|x - g^\star_{k_1+1}\|^2) - (\|x - g^\star_{k_1+1}\|^2 - \|x - g^\star_{k_1+2}\|^2) = \sigma_{k_1+1}(R)^2 - \sigma_{k_1+2}(R)^2 \geq 0,$$

which is exactly discrete convexity of the squared optimal error in k_1. □

Remark 3. *The proof above is stated for the residual low-rank interpretation of 2D GA-Planes: the low-resolution plane component is fixed before analyzing the rank-k_1 line component. This is the clean setting in which the Eckart-Young theorem gives an exact expression for the optimal error curve. For a fully joint nonconvex optimization over the low-resolution plane component and the low-rank line component at every k_1, the optimal plane component may in principle depend on k_1. In that more general setting, the same conclusion will require additional conditions to ensure that the optimal low-resolution component does not change in a way that violates the diminishing-returns behavior.*

Lemma GAP.3 (GA-Planes satisfies Assumption A3)**.** *There exists a finite $k_1^\star$ such that*

$$\|x - g^\star_{k_1^\star}\| = 0.$$

Proof. Fix $r_1 = N$ so that the linear interpolation at integer coordinates returns the stored grid values exactly. Let $r = \operatorname{rank}(M) \leq N$ and let $M = U\Sigma V^\top$ be the singular value decomposition with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. We exhibit a parameter setting at line feature dimension $k_1^\star = r$ that realizes M exactly. It follows from $r_1 = N$ that $e_1[i] = g_1[i, :]$ and $e_2[j] = g_2[j, :]$ for every (i, j). Set

$$g_1[i, l] = \sqrt{\sigma_l} U[i, l], \qquad g_2[j, l] = \sqrt{\sigma_l} V[j, l], \qquad W_{\text{line}}[l] = 1,$$

for $l = 1, \ldots, r$. Zero out the plane and bias components $g_{12} = 0, W_{\text{plane}} = 0, b = 0$. Substituting into the model output gives, for every (i, j),

$$\begin{aligned} g_d(\theta) &= W_{\text{line}}^\top (e_1 \circ e_2) + W_{\text{plane}}^\top e_{12} + b \\ &= \sum_{l=1}^{k_1} 1 \cdot g_1[i, l] \cdot g_2[j, l] \\ &= \sum_{l=1}^{r} \sigma_l U[i, l] V[j, l] = M[i, j], \end{aligned}$$

which is the SVD reconstruction of M. Hence, the reconstruction error vanishes at $k_1^\star = r \leq N$ with $r_1 = N$ and any arbitrary r_2, k_2. □

C Extensions of the Global Error Bound

C.1 Tightness of the $(d^\star - d)$ factor

The bounds of Section 3 are valid but, as Figures 1 and 2 show, conservative by orders of magnitude. Here, we show that the factor $(d^\star - d)$ is unimprovable under Assumptions A1–A3 alone, and that quantifying the rate of diminishing returns (tightening A2 to geometric error decrease) removes the $(d^\star - d)$ factor entirely and provides a mathematical foundation for our calibrated error estimates.

Theorem 8 (Worst-case tightness of the factor $(d^\star - d)$)**.** *Consider any finite parameter budget $d < d^\star$. Under Assumptions A1–A3, the factor $(d^\star - d)$ in Theorem 1 is worst-case tight: no uniformly smaller multiplier can be guaranteed from these assumptions alone.*

Proof. For tightness, fix any unit vector u and any $\Delta > 0$, and define $g_j^\star = x - \max\{0, d^\star - j\}\Delta u$ for any $j \geq 0$. Let $e_j := \|x - g_j^\star\|$. Then,

$$e_j = (d^\star - j)^+ \Delta.$$

It follows that Assumption A1 holds and Assumption A2 is also satisfied because the marginal error reductions equal Δ for $j < d^\star$ and zero thereafter. $e_{d^\star} = 0$ so Assumption A3 also holds. Moreover, for every $j < d^\star$,

$$\|g_j^\star - g_{j+1}^\star\| = \Delta \qquad \text{and} \qquad \|x - g_j^\star\| = (d^\star - j)\Delta.$$

Thus, Theorem 1 holds with equality at every $j < d^\star$; no scaling smaller than $(d^\star - d)$ is valid under Assumptions A1–A3 alone. □

The constructed instance is exactly the one in which no diminishing returns are available: Assumption A2 holds with equality at every increment, so the optimal error decreases linearly in d and the telescoping step of Theorem 1 loses nothing. However, Theorem 8 does not preclude improvements for specific architectures whose curves are known to decay faster, and the following theorem gives one such improvement.

Theorem 9 (A tighter bound under geometric diminishing returns)**.** *Let $e_j := \|x - g_j^\star\|$ and $\delta_j := e_j - e_{j+1}$. Suppose Assumptions A1 and A3 hold, and that Assumption A2 is strengthened to a geometric decay condition requiring $\delta_{j+1} \leq \rho\, \delta_j$ for $j = d, \ldots, d^\star - 2$ and some $\rho \in [0, 1)$. Then*

$$\|x - g_d^\star\| \leq \frac{1 - \rho^{d^\star - d}}{1 - \rho} \|g_d^\star - g_{d+1}^\star\| \leq \min\left\{ d^\star - d, \frac{1}{1 - \rho} \right\} \|g_d^\star - g_{d+1}^\star\|.$$

Proof. Assumption A3 and telescoping give $e_d = \sum_{i=0}^{d^\star - d - 1} \delta_{d+i}$. The geometric-decay condition implies $\delta_{d+i} \leq \rho^i \delta_d$, so

$$e_d \leq \delta_d \sum_{i=0}^{d^\star - d - 1} \rho^i = \frac{1 - \rho^{d^\star - d}}{1 - \rho} \delta_d.$$

Finally, $\delta_d \leq \|g_d^\star - g_{d+1}^\star\|$ by triangle inequality. The second inequality follows from $\sum_{i=0}^{d^\star - d - 1} \rho^i \leq d^\star - d$ and $\sum_{i=0}^{d^\star - d - 1} \rho^i \leq (1 - \rho)^{-1}$. □

Remark 4 (Interpretation of the calibration constant c)**.** *Theorem 9 replaces the budget-dependent factor $(d^\star - d)$ with the budget-independent constant $(1 - \rho)^{-1}$, which is determined by the architecture's rate of diminishing returns ρ rather than by the signal or the compression level. This offers a meaningful interpretation of the single per-architecture constant c calibrated in Section 4.1: c plays the role of an effective $(1 - \rho)^{-1}$, and its observed transferability across signals, dimensionalities, and compression levels is exactly what a decay rate fixed by the architecture would predict. We emphasize that the correspondence is interpretive rather than exact, since c is fit empirically and also absorbs the optimization-gap terms of Theorem 2 and the envelope width $U(d) - L(d)$ for architectures where Theorem 3 applies.*

C.2 Relaxing the exact-fit assumption

Assumption A3 requires that the parameterization represent x exactly at some finite budget $d^\star$, which can often be large in theory and in practice. For architectures outside the families we study, exact representation may be unavailable or difficult to certify. Since the assumption only enters our proofs to anchor a telescoping sum at a budget where the residual error is known, any budget with a known small residual serves the same purpose. Consider the following relaxation of Assumption A3.

Assumption A3′ (Approximate zero error). For some $\epsilon \geq 0$, there exists a finite budget $d^\star(\epsilon)$ with

$$\|x - g_{d^\star(\epsilon)}^\star\| \leq \epsilon. \tag{A3′}$$

We take $d^\star(\epsilon) := \min\{\, d \geq 0 : \|x - g_d^\star\| \leq \epsilon \,\}$ whenever this set is nonempty. Under Assumption A1, $d^\star(\epsilon)$ is nonincreasing in ϵ, and under Assumption A3 we have $d^\star(0) = d^\star$, so A3′ is implied by A3 for every $\epsilon \geq 0$ with $d^\star(\epsilon) \leq d^\star$.

Theorem 10 (Global error bound under approximate zero error)**.** *Under Assumptions A1, A2, and A3′, for any parameter budget $d \le d^\star(\epsilon)$, the compression error is bounded by*

$$\|x - g_d^\star\| \le (d^\star(\epsilon) - d)\|g_d^\star - g_{d+1}^\star\| + \epsilon. \tag{8}$$

Moreover, for any converged model outputs $\tilde{g}_d, \tilde{g}_{d+1}$ with optimization errors η_d, η_{d+1},

$$\|x - \tilde{g}_d\| \le (d^\star(\epsilon) - d)\|\tilde{g}_d - \tilde{g}_{d+1}\| + (d^\star(\epsilon) - d)\|\eta_{d+1} - \eta_d\| + \|\eta_d\| + \epsilon. \tag{9}$$

Setting $d^\star(\epsilon) = d^\star$ recovers the original results in Theorems 1 and 2.

Proof. Telescoping to $d^\star(\epsilon)$ gives

$$\|x - g_d^\star\| = \|x - g_{d^\star(\epsilon)}\| + \sum_{i=1}^{d^\star(\epsilon)-d} (\|x - g_{d+i-1}^\star\| - \|x - g_{d+i}^\star\|).$$

By Assumption A3′, $\|x - g_{d^\star(\epsilon)}\| \le \epsilon$, and by Assumption A2, every summand is at most $\|x - g_d^\star\| - \|x - g_{d+1}^\star\|$. Hence

$$\|x - g_d^\star\| \le (d^\star(\epsilon) - d)(\|x - g_d^\star\| - \|x - g_{d+1}^\star\|) + \epsilon.$$

Triangle inequality gives

$$\|x - g_d^\star\| - \|x - g_{d+1}^\star\| \le \|g_d^\star - g_{d+1}^\star\|,$$

which proves the first claim.

For the trained model outputs, triangle inequality yields

$$\|x - \tilde{g}_d\| \le \|x - g_d^\star\| + \|\eta_d\|.$$

In addition,

$$\|g_d^\star - g_{d+1}^\star\| \le \|\tilde{g}_d - \tilde{g}_{d+1}\| + \|\eta_{d+1} - \eta_d\|.$$

Substituting these inequalities into the first bound proves the second bound. □

Assumption A3′ asks only that the model approximately represent x to accuracy ϵ at some finite parameter budget, which follows from standard universal approximation or approximation rate statements and does not require the exact interpolation constructions of Lemmas G.3, F.4, I.4, and GAP.3. It is therefore easier to certify for a new architecture and robust to factors that may make exact representation unachievable in practice, such as finite-precision arithmetic or quantized parameters. Moreover, since $d^\star(\epsilon) \le d^\star$, this relaxation trades an additive ϵ against a smaller multiplier, potentially yielding tighter error bounds.

C.3 Batched Global Error Bound

In practice, estimating $\|g_d^\star - g_{d+1}^\star\|$ from a single adjacent-size pair of models can be sensitive to training instabilities. The following theorem addresses this by averaging the measured gap over a contiguous block of b compression increments, at the cost of b additional model fits.

Theorem 11 (Batched Global Error Bound)**.** *Under Assumptions A1 - A3, for any parameter budget $d \le d^\star$ and any block size $1 \le b \le d^\star - d$, the compression error is bounded by*

$$\|x - g_d^\star\| \le (d^\star - d) \cdot \frac{1}{b} \sum_{i=1}^{b} \|g_{d+i-1}^\star - g_{d+i}^\star\|. \tag{10}$$

Proof. We prove the batched version by averaging the one-step decreases of the optimal error curve over the first block of size b, and then using discrete convexity to show that every subsequent block has no larger average decrease. Define the one-step improvement

$$\delta_i := \|x - g_i^\star\| - \|x - g_{i+1}^\star\|, \qquad i = d, d+1, \ldots, d^\star - 1.$$

By Assumption A1, every $\delta_i \ge 0$, and by Assumption A2, the sequence $\delta_d \ge \delta_{d+1} \ge \cdots \ge \delta_{d^\star-1} \ge 0$ is non-increasing. Since $\|x - g_{d^\star}^\star\| = 0$ by Assumption A3, the error telescopes as

$$\|x - g_d^\star\| = \sum_{i=1}^{d^\star-d} \delta_{d+i-1}. \tag{11}$$

We partition the $d^\star - d$ indices into $B = \lceil (d^\star - d)/b \rceil$ consecutive blocks: the first $B-1$ blocks each have size b, and the last block has size $b' = d^\star - d - (B-1)b \le b$. If $B = 1$, then the partition consists only of the first block, so the block-sum argument will reduce to the telescoping identity $\|x - g_d^\star\| = \sum_{i=1}^{b} \delta_{d+i-1}$, and the stated bound follows directly by applying the triangle inequality to each δ_{d+i-1}. We thus consider $B > 1$ and bound each block's contribution separately, using the sum of the first block as a common majorant.

Blocks $k = 1, \dots, B-1$ (size b). For the first block $k = 1$, the inequality $\sum_{i=1}^{b} \delta_{d+i-1} \le \sum_{i=1}^{b} \delta_{d+i-1}$ holds trivially. For each subsequent block $k \ge 2$, monotonicity of δ gives $\delta_{d+(k-1)b+i-1} \le \delta_{d+i-1}$ for each $i = 1, \dots, b$, since $(k-1)b \ge b \ge 1$. Summing over i,

$$\sum_{i=1}^{b} \delta_{d+(k-1)b+i-1} \le \sum_{i=1}^{b} \delta_{d+i-1}.$$

Last block (size $b' \le b$). Every index in this block satisfies $d + (B-1)b + i - 1 \ge d + b$ for $B \ge 2$, so each summand obeys $\delta_{d+(B-1)b+i-1} \le \delta_{d+b-1}$. Since δ_{d+b-1} is the minimum of the first block, it is at most the block average:

$$\sum_{i=1}^{b'} \delta_{d+(B-1)b+i-1} \le b'\, \delta_{d+b-1} \le b' \cdot \frac{1}{b} \sum_{i=1}^{b} \delta_{d+i-1}.$$

Substituting these bounds into Equation (11) yields

$$\begin{aligned}
\|x - g_d^\star\| &= \sum_{k=1}^{B-1} \sum_{i=1}^{b} \delta_{d+(k-1)b+i-1} + \sum_{i=1}^{b'} \delta_{d+(B-1)b+i-1} \\
&\le (B-1) \sum_{i=1}^{b} \delta_{d+i-1} + b' \cdot \frac{1}{b} \sum_{i=1}^{b} \delta_{d+i-1} \\
&= \big[(B-1)b + b'\big] \cdot \frac{1}{b} \sum_{i=1}^{b} \delta_{d+i-1} \\
&= (d^\star - d) \cdot \frac{1}{b} \sum_{i=1}^{b} \delta_{d+i-1},
\end{aligned}$$

where the last line follows from our construction $(B-1)b + b' = d^\star - d$. Finally, the triangle inequality gives $\delta_{d+i-1} = \|x - g_{d+i-1}^\star\| - \|x - g_{d+i}^\star\| \le \|g_{d+i-1}^\star - g_{d+i}^\star\|$ for each i, so

$$\|x - g_d^\star\| \le (d^\star - d) \cdot \frac{1}{b} \sum_{i=1}^{b} \|g_{d+i-1}^\star - g_{d+i}^\star\|. \qquad \square$$

Setting $b = 1$ recovers Theorem 1. Empirically, averaging over a small block of nearby model pairs yields smoother, tighter, and more reliable bound estimates while preserving the same local-to-global interpretation. We therefore use the single-step version as the core theorem and reserve the batched version as a practical refinement requiring training a few additional models at different sizes.

D Experimental Details

D.1 Test Signals

We evaluate our framework on a diverse suite of synthetic and real signals spanning 2D and 3D domains, drawn from the benchmark of Kim and Fridovich-Keil [52] and standard datasets in the computational imaging literature. All signals are either generated at or resized to resolution 128^2 for 2D and 64^3 for 3D, ensuring that the model parameter budgets used in our experiments remain in the compressive regime.

Table 1: **Fixed hyperparameters for 2D signals.** We use these settings for 2D direct fitting and 2D MRI experiments. For GA-Planes, we use a linear decoder, consistent with the theoretical analysis.

Model	Hyperparameter	Value
FFN	Learning Rate	1e-3
	Embedding Size	256
	Hidden Layers	1
	σ	20
Instant-NGP	Learning Rate	5e-4
	Hidden Layers	1 (2-layer MLP)
	L	5
	F	2
	Hidden Layer Width	64
GA-Planes	Learning Rate	5e-3
	Line Resolution	size per axis of signal
	Plane Resolution	2
Grid	Learning Rate	5e-2

Table 2: **Fixed hyperparameters for 3D signals.** We use these settings for 3D direct fitting and 3D MRI experiments. For GA-Planes, we use a linear decoder, consistent with the theoretical analysis.

Model	Hyperparameter	Value
FFN	Learning Rate	8e-4
	Embedding Size	1520
	Hidden Layers	1
	σ	5
Instant-NGP	Learning Rate	5e-4
	Hidden Layers	1 (2-layer MLP)
	L	4
	F	2
	Hidden Layer Width	64
GA-Planes	Learning Rate	8e-3
	Line Resolution	size per axis of signal
	Plane Resolution	size per axis of signal
	Volume Resolution	2
	Volume Feature Dimension	1
Grid	Learning Rate	5e-2

Synthetic signals (2D, 3D). We construct two families of 2D and 3D synthetic signals parameterized by an explicit notion of effective bandwidth, following Kim and Fridovich-Keil [52]. *Bandlimited* signals are generated by sampling random noise and applying radial low-pass filters with exponentially spaced cutoffs in the Fourier domain. This yields signals whose frequency content, and thus approximation difficulty, can be smoothly varied from low to high bandwidth. *Sphere* signals consist of randomly placed filled disks (2D) or balls (3D) whose complexity is governed by the number and scale of the primitives: fewer larger objects produce "low-bandwidth" signals, while many small objects produce fine-grained "high-bandwidth" structure. We use a fixed intermediate bandwidth for both of these signals for all synthetic-signal experiments.

Real signals. To complement the controlled synthetic setting, we evaluate on representative real-world data spanning natural images and volumetric signals. For direct signal fitting, we use resized high-resolution images from the DIV2K dataset [48] and 3D Stanford Dragon surface volume [49]. For inverse problems, we consider two settings with qualitatively different forward models. In MRI reconstruction, we use brain volumes from the ATLAS stroke lesion dataset [50], which constitutes an underdetermined linear inverse problem. In radiance field reconstruction, we use the Lego scene from the NeRF-Blender synthetic dataset [51], where the forward model is the nonlinear volume rendering integral [51, 59] and supervision is provided through a photometric loss on posed training images. Together, these datasets introduce broadband spectral content, sharp geometric boundaries, and nontrivial spatial correlations that are not present in synthetic signals but are critical for assessing the performance of our framework beyond idealized settings.

Table 3: **Fixed hyperparameters for radiance fields.** For radiance field reconstruction, we depart from the shallow decoder theoretical setting by using deeper nonlinear MLP decoders for FFN, Instant-NGP, and GA-Planes. We also use total variation regularization for Grid.

Model	Hyperparameter	Value
FFN	Learning Rate	5e-4
	Embedding Size	512
	Hidden Layers	4
	σ	8
Instant-NGP	Learning Rate	5e-4
	Hidden Layers	2 (3-layer MLP)
	L	13
	F	2
	Hidden Layer Width	64
GA-Planes	Learning Rate	5e-4
	Line Resolution	200
	Plane Resolution	16
	Volume Resolution	8
	Volume Feature Dimension	4
	Hidden Layer Width	128
	Hidden Layers	2
Grid	Learning Rate	5e-2
	Total Variation Weight	1e-5

Table 4: **Calibrated constants used for predicted error estimates.** For each architecture, the constant c rescales the adjacent-model difference $B(d)$ into the practical predictor $cB(d)$, as described in Section 4.1. We use a separate constant for radiance fields ("NeRF") because radiance-field reconstruction uses a different rendering pipeline, model-size scale, and view-dependent output structure from the direct fitting and MRI experiments. Within each group of tasks, c is fixed across all datasets, signals, and compression levels.

Model	Task	Calibration constant c
FFN	NeRF	3.9
	All other tasks	1.4
Instant-NGP	NeRF	2.2
	All other tasks	0.9
GA-Planes	NeRF	4.4
	All other tasks	4.6
Grid	NeRF	3.1
	All other tasks	4.3

Design rationale. This combination of synthetic and real signals, and direct and indirect training supervision, provides a structured and realistic testbed for validating both the theoretical predictions and empirical behavior of our framework. It ensures that the observed trends, such as the scaling of reconstruction error with model size and the tightness of adjacent-model error predictors, are not artifacts of a specific data regime but instead reflect general properties of differentiable compressive parameterizations.

D.2 Implementation Details

To construct the global error curves, we vary one primary capacity parameter for each architecture while holding all other hyperparameters fixed. Specifically, we vary grid resolution for Grid, MLP width for FFN, hash table size for Instant-NGP, and feature dimension for GA-Planes. The fixed hyperparameters are reported in Table 1 for 2D direct fitting and 2D MRI, Table 2 for 3D direct fitting and 3D MRI, and Table 3 for radiance field reconstruction. For 2D and 3D signal fitting and MRI, we use shallow MLP decoders consistent with the theoretical setting. For radiance field reconstruction, we use deeper nonlinear decoders with view-direction inputs, following the standard NeRF setup [51], to improve rendering quality in the more challenging view-synthesis setting. All non-radiance-field experiments are implemented in PyTorch while radiance-field experiments are implemented in JAX.

For direct signal fitting, models are optimized against the observed signal values in image or volume space. For MRI reconstruction, models map spatial coordinates to image intensities and are supervised

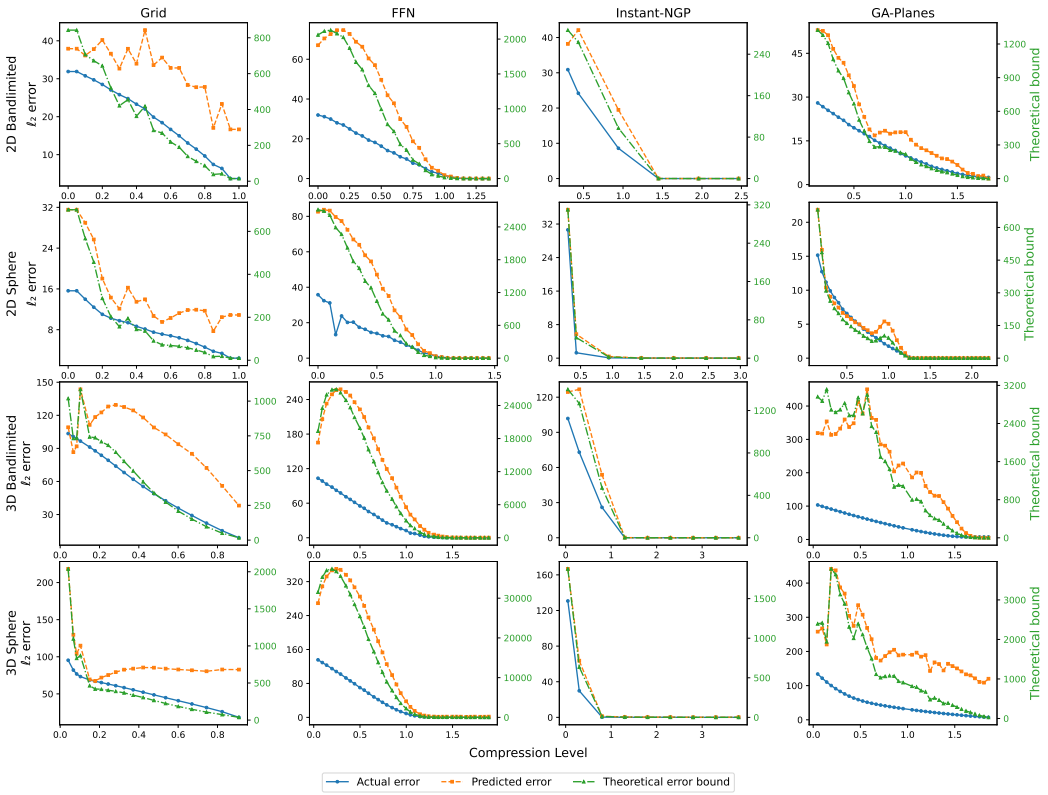

Figure 6: **Global error curves for direct fitting of synthetic signals.** We fit Grid, FFN, Instant-NGP, and GA-Planes directly to 2D and 3D synthetic signals, including bandlimited and sphere targets. The actual error (blue, left axis) is the ℓ_2 distance between reconstruction and ground truth, the optimization-gap-adjusted theory-dervied bound (green, right axis) follows from Theorem 2 as a function of compression level, and the predicted error (orange, left axis) rescales this bound by the constant c applied to the per-increment quantity $B(d)$ (Section 4.1). Although the theoretical bounds are conservative by orders of magnitude, the calibrated predictions track the actual error across architectures, signal types, and dimensionalities.

through a Fourier-domain k-space loss. For radiance-field reconstruction, models are trained from posed images using the standard photometric rendering loss. In all experiments, predicted global errors are computed from adjacent trained models using the calibrated architecture-specific constant described in Section 4.1, and local error heatmaps use the same constant applied pointwise to the difference between small- and large-model reconstructions. For radiance fields, the global error bounds are averaged over 10 test views while the model outputs used to compute the local error heatmaps are rendered from a single held-out test view, matching the evaluation procedure for the actual reconstruction error. We note that since the radiance field experiments use deeper, view-dependent decoders that lie outside our shallow-decoder theoretical setting, we calibrate c separately for radiance fields and for all remaining tasks. Table 4 reports the resulting constants, fit once per architecture per setting following Section 4.1 and held fixed across all signals, datasets, and compression levels.

All models are trained with the Adam optimizer using $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For optimization gap estimation, we require at least 5 random restarts before considering the restart plateau, and we use a tolerance of 10^{-6} for the change in the maximum pairwise reconstruction distance across restarts. Fourier feature networks use Gaussian-initialized Fourier embeddings, following prior work [11]. For the radiance field experiments, we additionally apply total variation regularization to the Grid baseline to encourage spatial smoothness and improve reconstruction quality, similar to Fridovich-Keil et al. [4]. All experiments were conducted on an NVIDIA RTX A6000 GPU with 48GB VRAM; memory usage did not impose a limiting constraint in our experiments.

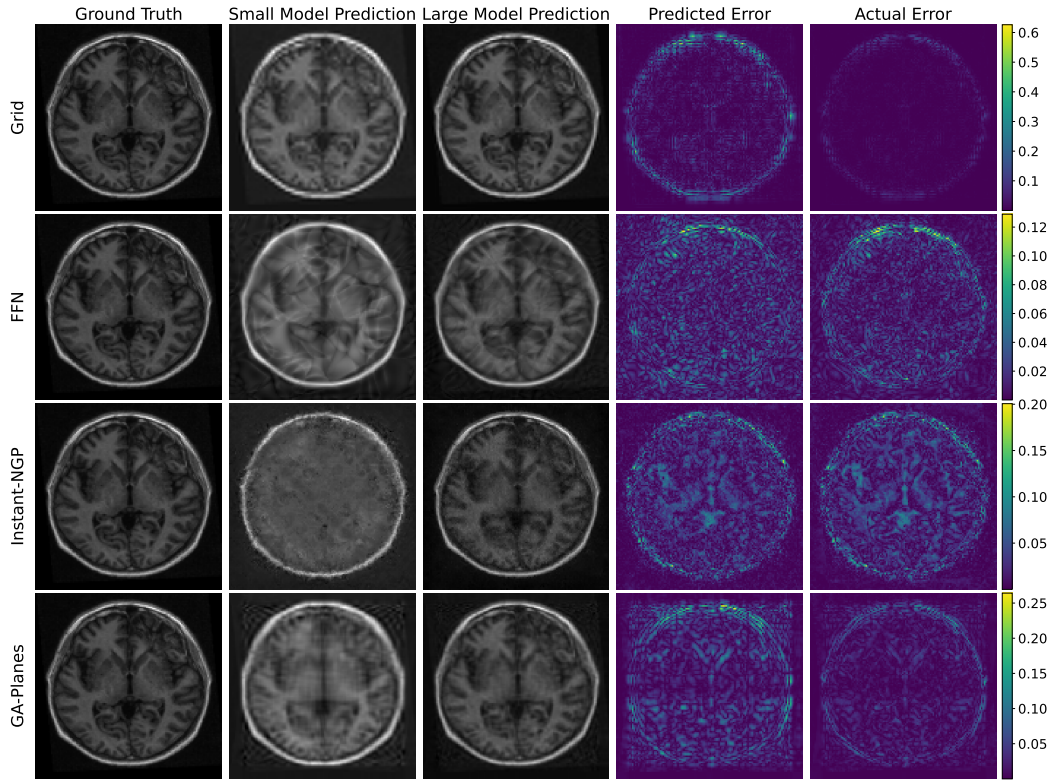

Figure 7: **Local error heatmaps for 2D MRI slice reconstruction.** Each row corresponds to a different architecture, with small and large models at compression levels $d/n \approx 0.2$ and 0.4, respectively. Each model maps 2D pixel coordinates to signal values and is supervised indirectly through a k-space loss on a single axial slice from the ATLAS dataset [50]. Across all four representations, the predicted error heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ captures the spatial structure of the actual error heatmap $|\tilde{g}_{\text{small}} - x|$, indicating that reconstruction artifacts can be localized without ground-truth access in the inverse-problem setting.

D.3 Global Error Curves: Extended Results

We provide additional global error curves for inverse problems, including 2D and 3D MRI reconstruction from k-space measurements and radiance field reconstruction from posed images, in Figure 2. Across these tasks, our calibrated predictor tracks the overall decay of the true reconstruction error despite indirect supervision. We also extend the global error evaluation to direct fitting of synthetic signals (from Kim and Fridovich-Keil [52]) in Figure 6. For each synthetic signal, we vary the model parameter budget while fixing the complexity parameter (effective bandwidth) at 0.3 for sphere signals and 0.9 for bandlimited signals. Across 2D and 3D bandlimited and sphere targets, the raw optimization-gap-adjusted theoretical bound is conservative, but the calibrated adjacent-model predictor follows the decay of the true reconstruction error. These results support the trend observed in the main text: adjacent-model differences provide a stable proxy for global compression error across architectures, signals, and dimensionalities.

D.4 Local Error Heatmaps: Extended Results and Quantitative Evaluation

We present additional local error heatmaps for two settings: 2D MRI reconstruction from k-space measurements (Figure 7) and direct fitting of the 3D Stanford Dragon surface volume (Figure 8). Across architectures, the predicted heatmaps recover the main spatial structure of the actual error maps, suggesting that adjacent-model differences remain informative for localizing compression artifacts in both direct-fitting and inverse-problem settings.

We also quantify this consistency by computing the Pearson correlation coefficient between the predicted heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ and the actual error heatmap $|\tilde{g}_{\text{small}} - x|$ for each signal in Figures 3–5 and Figures 7–8. Since the practically relevant question is not only whether the maps

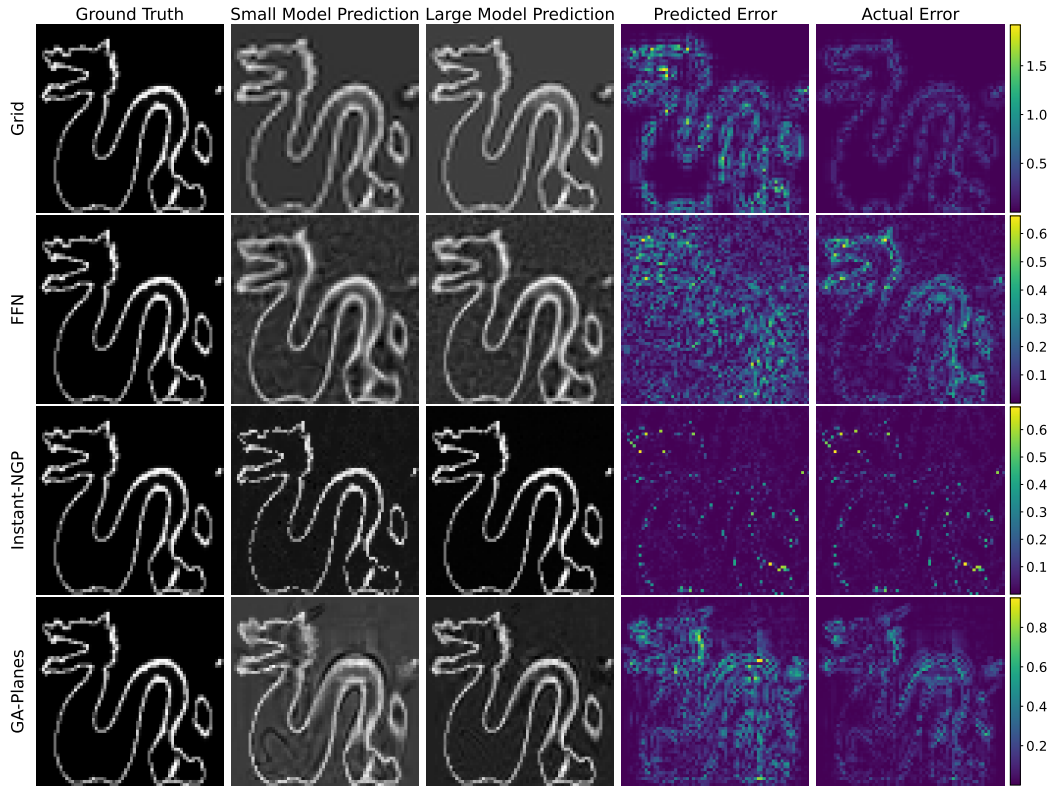

Figure 8: **Local error heatmaps for direct fitting of 3D Stanford Dragon surface volume [49].** Each row corresponds to a different architecture, with small and large models at compression levels $d/n \approx 0.2$ and 0.4, respectively. Across all four representations, the predicted error heatmap $c|\tilde{g}_{\text{small}} - \tilde{g}_{\text{large}}|$ captures the spatial structure of the actual error heatmap $|\tilde{g}_{\text{small}} - x|$, showing that local compression artifacts can be estimated without ground-truth access for 3D volumes.

agree but at what spatial resolution, we average-pool both maps into non-overlapping $p \times p$ patches before correlating, sweeping $p \in \{1, 2, 4, 8\}$ with $p = 1$ corresponding to the original pixel resolution. Table 5 reports results for the single signal shown in the corresponding figure, not averages across multiple signals. For every architecture and displayed signal, correlation increases monotonically with patch size. Pixel-level correlations range from 0.624 to 0.782 while (8×8)-patch correlations range from 0.841 to 0.981. The predicted heatmaps therefore show substantial spatial agreement with the actual errors even at pixel resolution, and stronger agreement at the patch or region-of-interest scale most relevant in practice.

Interpretation and scope. Since Pearson correlation is invariant to positive scaling, the constant c does not affect these values. The experiment therefore evaluates the localization of the adjacent-model difference itself rather than the calibration of its magnitude. Moreover, these single-example results measure empirical spatial correlation and do not establish a pointwise, region-wise, or distribution-free error guarantee. These results serve to quantify the visual agreement between the predicted and actual error maps across different signals and inverse problems.

D.5 Model-size Selection without Ground Truth

We evaluate whether our calibrated predictor can support a concrete downstream decision: given an error tolerance τ, select the smallest parameter budget whose MRI reconstruction satisfies

$$\|x - \tilde{g}_d\|_2 \leq \tau,$$

without observing the test signal x. We use an FFN [11] model and sweep its capacity over ten compression levels $d/n \in \{0.1, 0.2, \ldots, 1.0\}$, using the hyperparameters from Table 1. For each test MRI slice, we fit the image-space FFN indirectly from undersampled k-space measurements, as in our other 2D MRI experiments. Our predicted error is $\hat{e}_{\text{adj}}(d) = cB(d)$, using the fixed FFN constant c from Table 4.

Table 5: **Pearson correlation between predicted and actual error heatmaps.** Both heatmaps are average-pooled over non-overlapping $p \times p$ patches before computing correlation, where $p = 1$ is the original pixel resolution. Each block corresponds the single signal displayed in the corresponding figure; values are not averages over multiple signals.

Model	$p = 1$	$p = 2$	$p = 4$	$p = 8$
DIV2K direct fitting (Figure 3)				
Grid	0.735	0.886	0.947	0.981
FFN	0.624	0.767	0.881	0.974
Instant-NGP	0.761	0.861	0.929	0.961
GA-Planes	0.782	0.877	0.935	0.965
3D MRI reconstruction from k-space(Figure 4)				
Grid	0.625	0.805	0.854	0.938
FFN	0.666	0.719	0.776	0.841
Instant-NGP	0.700	0.790	0.897	0.930
GA-Planes	0.750	0.866	0.926	0.954
NeRF radiance-field reconstruction (Figure 5)				
Grid	0.749	0.806	0.862	0.911
FFN	0.647	0.731	0.796	0.845
Instant-NGP	0.781	0.869	0.919	0.949
GA-Planes	0.764	0.846	0.897	0.930
2D MRI reconstruction from k-space (Figure 7)				
Grid	0.780	0.910	0.958	0.977
FFN	0.601	0.710	0.798	0.871
Instant-NGP	0.870	0.927	0.960	0.975
GA-Planes	0.808	0.882	0.928	0.958
3D Stanford Dragon surface volume direct fitting (Figure 8)				
Grid	0.654	0.799	0.863	0.885
FFN	0.538	0.695	0.773	0.860
Instant-NGP	0.988	0.992	0.996	0.998
GA-Planes	0.761	0.861	0.807	0.933

A natural ground-truth-free heuristic for this task is to transfer the error statistics measured on a known dataset. We compute the mean reconstruction error

$$\hat{e}_{\mathrm{trans}}(d) = \frac{1}{200} \sum_{i=1}^{200} \|x_i - \tilde{g}_{d,i}\|_2$$

and its empirical standard deviation $\hat{\sigma}(d)$ for FFN at each compression level over 200 central slices from the Brain subset of FastMRI [60]. We evaluate two transferred baselines: the mean error $\hat{e}_{\mathrm{trans}}(d)$ and the more conservative mean-plus-standard-deviation estimate

$$\hat{e}_{\mathrm{trans+std}}(d) = \hat{e}_{\mathrm{trans}}(d) + \hat{\sigma}(d).$$

Both estimates are fixed across two test signals: a held-out fastMRI brain slice not in the reference set (in-distribution) and an ATLAS [50] slice (out-of-distribution). The baseline statistical approaches and our method are ground-truth-free at test time but use offline ground-truth calibration: the baselines transfer an entire mean-error curve, whereas our method transfers one architecture-specific scalar c and recomputes the adjacent-model difference using the new measurements. We report results in Table 6, with actual errors included only for retrospective evaluation.

Given an estimated error curve $\hat{e}$, we select $\hat{d}_\tau = \min\{d : \hat{e}(d) \leq \tau\}$ over the tested budgets. For $\tau = 15$, both baselines select $d/n = 0.2$ for both test slices: the mean and mean-plus-standard-deviation error estimates at this budget are 11.71 and 13.78, respectively. This choice satisfies the

Table 6: **Ground-truth-free error estimation for 2D MRI.** The transferred baseline is the mean FFN error over 200 central slices from the Brain subset of FastMRI [60]; "Ours" is the calibrated adjacent-model estimate $cB(d)$. Actual errors are shown only for evaluation and are unavailable to either selection rule at test time. All entries use the same ℓ_2 norm convention.

	Transferred FastMRI	Held-out FastMRI		ATLAS	
d/n	Mean $\pm$ std	Actual	Ours	Actual	Ours
0.1	17.24 ± 2.99	17.14	35.25	18.74	39.41
0.2	11.71 ± 2.07	14.27	29.79	17.07	35.16
0.3	9.07 ± 1.72	11.22	21.30	16.02	35.99
0.4	8.31 ± 1.52	10.15	19.07	15.03	29.57
0.5	7.07 ± 1.22	8.34	14.19	14.26	26.81
0.6	5.76 ± 0.99	6.77	11.16	13.72	25.19
0.7	4.30 ± 0.73	5.06	9.95	12.09	20.39
0.8	3.07 ± 0.48	3.53	7.97	9.49	19.79
0.9	1.39 ± 0.22	1.56	3.30	7.26	12.66
1.0	0.54 ± 0.09	0.56	1.57	3.87	6.77

target on the held-out FastMRI slice, whose actual error is 14.27, but fails on the ATLAS slice, whose actual error is 17.07. Our calibrated adjacent-model predictor instead selects $d/n = 0.5$ and 0.9, respectively, with corresponding actual errors of 8.34 and 7.26, both satisfying the target. Even adding two standard deviations to the transferred mean selects $d/n = 0.3$ on ATLAS, whose actual error of 16.02 still exceeds the target $\tau = 15$.

This is precisely the failure mode that motivates a signal-adaptive, theory-driven predictor over a transferred statistical estimator. At the same compression level $d/n = 0.5$, the transferred baseline assigns both slices the predicted error 7.07, whereas our estimate increases from 14.19 on FastMRI to 26.81 on ATLAS, mirroring the increase in actual error from 8.34 to 14.26. Across all ten compression budgets and both test slices, our error estimate remains a valid upper bound in all 20 cases, compared with one case for the transferred mean and five for the mean-plus-standard-deviation baseline. Note that these counts describe empirical coverage on this illustrative two-slice comparison, not a distribution-free guarantee.

D.6 Calibration Sensitivity and Predictor Tightness

We quantify the stability of the calibration constant c and the tightness of the resulting predictor over 30 signals: 20 DIV2K images under direct fitting and 10 ATLAS MRI volumes reconstructed from k-space measurements. For each architecture, we evaluate ten compression ratios $d/n \in \{0.1, 0.2, \ldots, 1.0\}$.

For signal i, let

$$e_i(d) = \|x_i - \tilde{g}_{i,d}\|_2, \qquad \hat{e}_i(d; c) \coloneqq cB_i(d),$$

where $B_i(d)$ is the optimization-gap-adjusted per-increment quantity from Section 4.1. We report the prediction ratio

$$R_i(d; c) \coloneqq \frac{\hat{e}_i(d; c)}{e_i(d)},$$

where $R = 1$ corresponds to exact error prediction, $R > 1$ is conservative but valid, and $R < 1$ is an underestimate (invalid error bound). Over the tested compression levels for which $B_i(d) > 0$, the smallest constant dominating the observed error on signal i is

$$c_i = \max_d \frac{e_i(d)}{B_i(d)}.$$

Sensitivity of c. We independently fit c_i on each of the 30 signals and evaluate the resulting predictor on the remaining signals and compression levels. To stress-test the calibration, we define $c_{\text{worst}} \coloneqq \min_i c_i$, the least conservative of the 30 fitted values. Table 7 reports the distribution of c_i across these 30 signals, together with the prediction ratio and empirical coverage obtained using

c_{worst}. Coverage is the fraction of held-out (signal, capacity) pairs for which $R \geq 1$ (our prediction is a valid upper bound on the true error).

Table 7: **Sensitivity of the calibration constant.** Each c_i is fit on a single signal. c_{worst} is the least conservative of the 30 values of c_i across all 30 signals, i.e. the choice most likely to underestimate compression error. Prediction ratios and coverage are evaluated on held-out (signal, capacity) pairs using c_{worst}. Coverage is empirical and does not constitute a distribution-free guarantee.

Model	Range of c_i	c_i, mean $\pm$ std	R under c_{worst}	Coverage under c_{worst}
Grid	$[4.03,\ 4.34]$	4.20 ± 0.09	2.08 ± 0.95	93.87%
FFN	$[1.35,\ 1.65]$	1.48 ± 0.10	2.32 ± 0.38	94.76%
Instant-NGP	$[0.63,\ 1.12]$	0.81 ± 0.12	1.97 ± 0.55	95.08%
GA-Planes	$[4.52,\ 4.83]$	4.63 ± 0.08	2.29 ± 0.36	94.38%

The coefficients of variation of c_i are 2.1%, 6.8%, 14.8%, and 1.7% for Grid, FFN, Instant-NGP, and GA-Planes, respectively. Thus, c is particularly stable for Grid and GA-Planes, moderately variable for FFN, and most signal-dependent for Instant-NGP. Nevertheless, even the least conservative calibration c_{worst} yields mean prediction ratios of 1.97–2.32 and empirical coverage of the true error on between 93.87% and 95.08% of (signal, capacity) pairs across both image-fitting and MRI reconstruction domains. Sub-100% coverage signifies exactly the occasional underestimates that the worst-case theoretical bounds of Section 3 exclude, at the cost of orders of magnitude of looseness on average.

Predictor tightness. We next use the fixed architecture-specific constants reported in Table 4, rather than refitting c for individual signals. Table 8 reports the mean and standard deviation of R, aggregated over all (signal, capacity) pairs within each dataset.

Table 8: **Predictor tightness across** 30 **signals.** Mean $\pm$ standard deviation of the predicted-to-actual error ratio R, using the fixed per-architecture calibration constants from Table 4. Statistics are aggregated over all (signal, capacity) pairs in each dataset. Values above 1 are conservative.

Dataset	Grid	FFN	Instant-NGP	GA-Planes
DIV2K	2.54 ± 1.16	2.31 ± 0.65	2.17 ± 0.66	2.26 ± 0.49
ATLAS MRI	1.96 ± 0.98	2.40 ± 0.35	1.91 ± 0.62	2.41 ± 0.64

Across four architectures and two datasets, the mean prediction ratio ranges from 1.91 to 2.54, so the approximate factor-of-two conservatism observed in Figures 1 and 2 reproduces across 30 signals rather than reflecting a favorable choice of example. Accuracy is comparable under the two forms of supervision: no architecture changes its mean ratio by more than 0.6 between direct fitting on DIV2K and k-space supervision on ATLAS, and neither dataset is systematically easier, with Grid and Instant-NGP predicting slightly tighter bounds on MRI while FFN and GA-Planes yield slightly tighter bounds on images. Grid has the largest variance on both datasets, with standard deviations of 1.16 on DIV2K and 0.98 on ATLAS. This is consistent with Grid's Fourier-shell characterization: its approximation error and adjacent-model gap depend directly on the signal's spectral decay, and therefore vary more with signal content than the other architectures.

Scope. We note that these results quantify empirical transfer and tightness on the tested signals; they do not convert the calibrated predictor $cB(d)$ into a certified bound. In particular, $R > 1$ on average does not imply $R \geq 1$ for every (signal, capacity) pair, and the reported coverage is not a population-level or distribution-free guarantee.

D.7 Computational Overhead

We measure sequential (fitting one model at a time) wall-clock time on an NVIDIA RTX A6000 GPU, normalized by the time required to train a single model at capacity d. We use the same parameter budget sweep of 20 DIV2K images and 10 ATLAS MRI volumes described in Appendix D.6. All models are trained independently without warm starts; results are shown in Table 9.

Our unadjusted adjacent-model predictor trains an additional model at the neighboring capacity $d+1$ and evaluates $c\|\tilde{g}_d - \tilde{g}_{d+1}\|_2$. Because this predictor requires one additional model fit, its total runtime is $1.99\times$–$2.01\times$ that of training a single model; computing the norm between two trained outputs adds negligible time.

Our restart-adjusted predictor additionally estimates the optimization-gap using the restart-based procedure described in Section 4.2 and evaluates the full optimization-gap-adjusted quantity $cB(d)$. Estimating the optimization gap is more expensive, requiring $5.09\times$–$6.78\times$ the time of training a single model since it requires independent restarts to estimate optimization gap at each capacity. The larger standard deviation reflects our adaptive stopping rule, under which the restart count is data-dependent rather than fixed. This cost is only incurred when evaluating the optimization-gap-adjusted expression, rather than by the basic adjacent-model calibrated predictor.

Table 9: **Computational overhead.** Mean $\pm$ standard deviation of wall-clock time, normalized by the runtime of one model fit at capacity d. The unadjusted predictor includes an independently trained neighboring model, while the restart-adjusted predictor additionally includes estimation of the optimization-gap terms η. We fit all models sequentially in these results.

Configuration	Grid	FFN	Instant-NGP	GA-Planes
Single fit	$1\times$	$1\times$	$1\times$	$1\times$
Unadjusted predictor	$2.01 \pm 0.14\times$	$2.00 \pm 0.04\times$	$1.99 \pm 0.08\times$	$2.01 \pm 0.08\times$
Restart-adjusted predictor	$5.09 \pm 1.99\times$	$6.78 \pm 1.64\times$	$6.46 \pm 1.80\times$	$6.47 \pm 1.75\times$

These figures are for a single parameter budget evaluated in isolation. In practice, the predictor is often used while sweeping consecutive capacities, which is standard for model selection with implicit neural representations and is the setting of Appendix D.5 and Appendix D.6. In this case, the cost can amortize because each trained model serves as the neighbor of its predecessor: M adjacent-model comparisons require $M+1$ model fits rather than $2M$. Independent fits and restarts are also directly parallelizable on multiple GPUs, which reduces latency but not total compute. The reported measurements are sequential and do not include either form of amortization.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our contributions are summarized at the end of the introduction and supported with theoretical and experimental results in both the main paper and the appendix. Proofs for all theoretical results in the main paper are included in Appendix B and C.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5 discusses the main limitations of our work, including the restriction of our theoretical guarantees to shallow MLP decoders and the fact that some empirical results do not yet have formal guarantees. We also outline future directions for addressing these limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper states the assumptions for the main theoretical results, and the full proofs are provided in the Appendix. For architecture-specific results, the corresponding assumptions and verification proofs are stated explicitly in the relevant appendix subsections.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The appendix provides the datasets, training configurations, and model parameters needed to reproduce the main experimental results supporting our claims. We have also released our code and included it as supplementary materials for reproducibility.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We release our code and include it as a public link. We use publicly available datasets and include scripts to generate the synthetic data.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4 introduces sufficient experimental details to understand the figures and the results in the paper. The appendix further specifies the training and evaluation details, including datasets and evaluation settings, model hyperparameters, the capacity parameter varied for each architecture, training losses, optimizer settings, restart protocol, and implementation details.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: We quantify the sensitivity of the calibration constant c and the tightness of the resulting predictor over 30 signals in Appendix D.6. We also account for optimization variability through the restart-based optimization-gap estimate described in Section 4.2 and Appendix D.2.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: As stated in Section D.2, all experiments were conducted on an NVIDIA RTX A6000 GPU with 48GB VRAM, with memory usage posing no limitations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper relies solely on available datasets and synthetic signals. No human subjects are involved, and we don't see any direct harmful applications. All processes and experiments conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses positive impacts through ground-truth-free error estimation for compressed signal representations, which could improve reliability in settings such as biomedical imaging, robotics, and bandwidth-limited sensing. It also discusses limitations and potential negative impacts: the calibrated predictors and local heatmaps are not formal certificates, so overinterpreting them in safety-critical applications could create false confidence. We emphasize that local predictions remain empirical and require further validation before deployment in high-stakes settings.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We rely on existing open-source code and generate fully synthetic data. We do not release any high-risk data or pretrained models. Our experiments use existing datasets and synthetic data; the ATLAS MRI data are used under the official access-controlled Terms of Use and are not redistributed. We do not process personally identifying information or release pretrained weights, so no additional safeguards beyond respecting dataset terms and standard open-source licensing are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite and credit all third-party assets used in our paper, and where applicable, we include their license types and URLs. For codebases: Fourier Feature Networks is released under the MIT License; GA-Planes under CC-BY 4.0; and Instant-NGP under NVIDIA Source Code License (non-commercial). For datasets: DIV2K is licensed under CC-BY 4.0. The Lego Blender model is under CC BY-NC 2.0; we use it only for non-commercial academic research and do not redistribute the original asset.

    We use the ATLAS stroke MRI dataset and cite the corresponding dataset descriptor and dataset source. The ATLAS dataset is distributed through an access-request process: users must agree to the ATLAS Terms of Use, complete a request form, and receive an encryption code; the code is not to be shared. We use the data only for research purposes, do not attempt to identify research subjects, do not redistribute the dataset or derived materials, and comply with the citation requirements specified by the ATLAS Terms of Use.

    We were unable to locate explicit licenses for the Stanford Dragon (https://graphics.stanford.edu/data/3Dscanrep/). In this case, we provide the original URL and cite its source appropriately. For synthetic signals, we provide applicable citations and clarify that these datasets are re-creations based on conceptual references, not direct copies of existing data files.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: No new datasets or external benchmark assets are introduced. We use existing public datasets and scenes. We provide documentation of the experimental setup, test signals, implementation details, and hyperparameters in Appendix D.2. Released code will include instructions for reproducing the main results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The study does not involve any crowdsourcing experiments or human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: IRB approvals are not applicable because no human participants are involved.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used only for typo checks and minor grammatical polishing. They did not contribute to the core methodology and theoretical results of the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.