SELECTING THE MINIMIZER: TEST-TIME DEPTH COMPLETION BEYOND THE MEASURED PIXELS

Anonymous authors
Paper under double-blind review

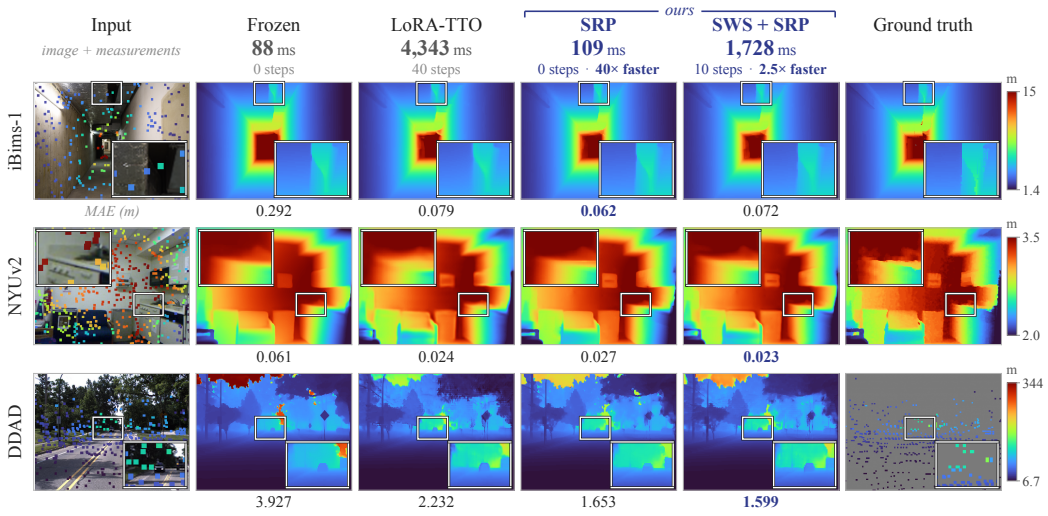

Figure 1: **Depth completion at three budgets: the reconstructions look alike, the cost does not.** Each head gives the latency at VGA on one RTX 4090 and the adaptation steps (speed-ups against LoRA-TTO), and each panel that sample's MAE (m), the lowest in bold. Rows are iBims-1, NYUv2 and DDAD, each **the first image of its benchmark**; the white box marks the window magnified in the corner. Depth scale is fixed per row (meters, right), and the input measurements are dilated for visibility. Accuracy over all 6,479 images is in Tab. 1.

ABSTRACT

Recent test-time optimization for depth completion adapts a frozen monocular metric-depth model to the sparse measurements of a single image through a low-rank adapter, and so needs no completion training set. Eliminating its scale-and-shift alignment analytically puts the objective in closed form, which explains why the procedure is fragile and slow: its minimizers leave at least 95% of the parameter dimensions free, so the difficulty is not reaching a minimizer but choosing one, and its standard zero initialization is degenerate, with one factor's gradient and curvature vanishing identically. We add one component for each. Semiparametric residual propagation (SRP) chooses the minimizer explicitly, in function space, splitting the correction into a parametric term in the predicted depth and a field harmonic on a bandwidth-free affinity graph, with the measurements imposed as constraints. A spectral warm start (SWS) replaces the degenerate initialization with the best rank-r descent step, whose amplitude is computed rather than searched. Across five benchmarks, 10 adaptation steps—a quarter of the baseline's budget—match or beat the full 40-step baseline on every metric while lowering MAE by 27% on KITTI and 31% on DDAD. SRP alone needs no backward pass, runs 30–40× faster than that adaptation and lowers its MAE by 22% on KITTI and 26% on DDAD. Code is available at https://anonymous.4open.science/r/selecting-the-minimizer-test-time-depth-completion/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Depth completion is normally supervised, and the resulting models inherit the sensor pattern, depth range and scene statistics of their training set. A recent alternative needs no completion training set: a frozen monocular metric-depth model (Piccinelli et al., 2026; Wang et al., 2025; Lin et al., 2026) already predicts a plausible dense surface, so at test time one fits a low-rank adapter (Hu et al., 2022) to a *single* image until the prediction agrees with that image's sparse measurements, and reads out the adapted dense prediction (Seo et al., 2026). Nothing is learned across images, so the same procedure applies to a lidar scan, a structured-light frame and a handful of visual-inertial points.

Low-rank test-time optimization (LoRA-TTO) is the cheapest route to depth completion—1.6–4.4$\times$ faster than prompt tuning (Jeong et al., 2025) across the resolutions we time (Tab. 7), and faster than guided diffusion (Viola et al., 2025) in the timings of Seo et al. (2026)—yet every image still costs tens of parameter updates, and the result depends on the seed. We argue that speed and stability are one problem, and that both follow from the structure of the test-time objective.

The structure is visible once the objective is written in closed form. Every test-time objective of this kind aligns the prediction to the measurements by scale and shift before taking the residual; that inner problem has an exact solution, and substituting it back (Sec. 3) reduces the objective to the variance of the measurements times one minus their squared correlation with the prediction. Three consequences follow. Every gradient carries the units of the data, so no step length stated in absolute terms can suit more than one benchmark. The standard zero initialization of the adapter is degenerate: the objective is constant in one factor there, so every run spends its first steps leaving that point in a direction chosen by the seed. And because the measurements cover at most a few percent of the pixels, the minimizers form a set retaining at least 95% of the parameter dimensions: the optimizer does not fail to find a minimizer, and *which* one it reaches—hence what the unmeasured pixels look like—is left to implicit bias.

The last two point to fixes, and by the first, neither fix carries a constant in data units. Semiparametric residual propagation (SRP, Sec. 4.1) makes the choice of minimizer explicit, in function space, where it is unique: it writes the correction to the aligned prediction as a parametric term in the predicted depth plus a field harmonic on an image-and-depth affinity graph, and imposes the measurements as constraints rather than penalties; the affinity is built from gradient ranks, so it has no bandwidth to tune. The spectral warm start (SWS, Sec. 4.2) replaces the degenerate initialization with the best rank-r descent step at the frozen weights, with its direction, per-layer scale and amplitude all computed in closed form. Because SRP is a read-out, it applies to whichever prediction reaches it—frozen, or adapted at any budget—so the two components trace one accuracy/cost curve (Fig. 1) rather than a single operating point.

Across five benchmarks and 6,479 images, no method we compare against lies below that curve on either metric (Sec. 5.3). With no backward pass at all, SRP on the frozen prediction is 30–40$\times$ faster end to end than the 40-step adaptation and lowers its MAE on both driving benchmarks; with 10 steps, a quarter of the baseline's budget, SWS + SRP matches or beats it on every benchmark and metric, lowering MAE by 27% on KITTI and 31% on DDAD. The analysis also predicts what should *not* help—better optimization and parameter-space regularization—and none of twelve such interventions improves on the baseline by more than half a percent (Sec. 5.4).

Contributions. **(i)** The closed form of the test-time objective and three exact consequences: a gradient scale set by the data, a degenerate standard initialization and a minimizer set of codimension at most 4.3% (Sec. 3; proofs in App. A). **(ii)** SRP, which chooses among those minimizers in function space, with the measurements as constraints, an affinity with no length scale and a parametric term that escapes the maximum principle bounding any propagated correction (Sec. 4.1). **(iii)** SWS, an analytic first step whose direction, per-layer scale and amplitude are all computed rather than searched (Sec. 4.2).

2 RELATED WORK

Depth completion without a completion training set. Recent monocular estimators predict metric depth across domains without per-dataset fine-tuning (Piccinelli et al., 2024; 2026; Yang et al., 2024); their output is geometrically consistent, but its absolute scale is unreliable—exactly the gap a

handful of measurements can close. Three lines exploit this without paired sparse and dense depth: test-time prompt tuning (Jeong et al., 2025), diffusion priors (Viola et al., 2025; Lee et al., 2025) and low-rank decoder adaptation (Seo et al., 2026); concurrent work (Ke et al., 2026) adapts a frozen 3D foundation model per sample in the same way. We study the last, the cheapest of the three and the one whose objective admits an exact characterization; the analysis applies to any method whose per-sample objective is an aligned masked regression against the measurements.

Adapter initialization. Among works that change how low-rank factors (Hu et al., 2022) are initialized or scaled (Meng et al., 2024; Hayou et al., 2024; Liu et al., 2024), closest to SWS is LoRA-GA (Wang et al., 2024), which also initializes from the singular vectors of the gradient at the frozen weights, as GaLore (Zhao et al., 2024) does for its projection subspace. Three differences are forced by our analysis: LoRA-GA subtracts its initial product so that the starting function is unchanged, whereas SWS takes the step; SWS scales each layer by its own Frobenius norm rather than by one absolute length; and its amplitude is computed, not tuned. Curvature spectra (Li et al., 2018) and sharpness-aware minimization (Foret et al., 2021) treat flat directions as a proxy for generalization over a dataset; here they are exact and known in closed form, and flatness-seeking does not help (Sec. 5.4).

Propagation and harmonic interpolation. Propagating sparse values along image affinities is classical (Levin et al., 2004; He et al., 2010), and supervised depth completion learns the affinities from paired data (Cheng et al., 2018; Park et al., 2020), which this setting rules out. The closest formulation comes from semi-supervised learning: Zhu et al. (2003) pose label propagation as a harmonic function on a weighted graph with the labeled points as Dirichlet boundary values. SRP departs from it in three ways. **(a)** Their affinity carries a length scale per feature, learned by entropy minimization; our rank affinity (5) has none. **(b)** Their field *is* the estimate, while ours corrects one—semiparametric, as in regression kriging (Hengl et al., 2007), but with a trend in the predicted depth rather than in position, so it is defined far from every measurement. **(c)** The maximum principle is a proof device there; here it is a design constraint whose effect we measure and whose limit the parametric term escapes.

3 What the Test-Time Objective Actually Is

Every statement in this section is exact: App. A proves each one, and App. E verifies them numerically and measures the size of the effects they force.

Setup. Let $\mathcal{M}$ be the set of m pixels of an image carrying a depth measurement, $y \in \mathbb{R}^m$ those measurements, and $f_\theta \in \mathbb{R}^N$ the dense prediction of the frozen model, whose decoder carries a low-rank adapter with parameters $\theta \in \mathbb{R}^D$ adding $\Delta W = s\,AB$ (scale s, rank r) at each adapted layer (App. C). Because monocular depth is defined only up to an affine map for many backbones, the objective aligns the prediction to the measurements before taking the residual:

$$\mathcal{L}(\theta) \;=\; \min_{a,b} \frac{1}{m} \sum_{i \in \mathcal{M}} \big(a\,[f_\theta]_i + b - y_i\big)^2. \tag{1}$$

Implementations recompute (a, b) at every step from the detached prediction and treat it as constant during backpropagation. This is not an approximation: (a, b) is the exact inner minimizer, so by the envelope theorem the detached gradient is the exact gradient of $\mathcal{L}$ (Prop. A.2).

3.1 The objective is a correlation criterion

Solving the inner least-squares problem and substituting back gives (Prop. A.1)

$$\mathcal{L}(\theta) \;=\; \mathrm{Var}_{\mathcal{M}}(y)\,\Big(1 - \rho_{\mathcal{M}}\big(f_\theta, y\big)^2\Big), \tag{2}$$

whenever $\mathrm{Var}_{\mathcal{M}}(f_\theta) > 0$. The identity is the classical coefficient of determination; our observation is that it *is* the test-time objective—the alignment is the standard scale-and-shift-invariant depth criterion (Ranftl et al., 2022)—so minimizing the aligned squared error is exactly maximizing the correlation between the prediction and the measurements on $\mathcal{M}$. Three consequences follow: (C1) and (C3) are the first two named in Sec. 1, and (C2) determines how curvature must be measured.

(C1) The gradient carries the units of the data. $\mathrm{Var}_{\mathcal{M}}(y)$ multiplies the whole objective, hence every gradient and every curvature eigenvalue. Across the benchmarks the loss at initialization spans nearly six orders of magnitude, while the dimensionless factor $1 - \rho^2_{\mathcal{M}}$ spans about two: most of the apparent difficulty gap between an indoor room and a highway is a change of units, and a step rule in absolute units cannot suit more than one benchmark. Adaptive optimizers mask this without removing it, spending steps to estimate from gradient magnitudes a factor that (2) gives in closed form.

(C2) The objective has two exact flat manifolds. Since (2) depends on f_θ only through a correlation, $\mathcal{L}(\alpha f_\theta + \beta) = \mathcal{L}(f_\theta)$ for any $\alpha \neq 0$ and β; separately, the low-rank product is invariant under $(A, B) \mapsto (AM, M^{-1}B)$ for *any* invertible $r \times r$ matrix M, an r^2-dimensional flat manifold per adapted layer (Prop. A.3). Motion along the affine manifold is wasted, since the alignment reabsorbs it, and curvature must be computed with the alignment differentiated, or the apparent sharpest direction is mostly a change of output scale (App. E).

(C3) The standard initialization is blind in one factor and indefinite in curvature. Low-rank adapters set $A = 0$ so that the adapted model starts at the frozen model. Since $\Delta W = s\,AB \equiv 0$ for *any* B when $A = 0$, the objective is constant in B there, not merely stationary: $\nabla_B \mathcal{L} \equiv 0$ and $\nabla^2_{BB} \mathcal{L} \equiv 0$. All seeds therefore start at the same point in function space and differ only in a direction the objective cannot see. Nor is the point a minimum: $\nabla_A \mathcal{L} = s\,(\nabla_{\Delta W} \mathcal{L}) B^\top$ does not vanish—it is the gradient SWS builds its step from—and with the BB block identically zero, the Hessian is indefinite whenever the frozen model's gradient is nonzero (Prop. A.4). A short run must grow A out of this point and still converge within the same budget.

3.2 The hard part is choosing a minimizer, not finding one

It is tempting to read the instability above as a hard nonconvex problem in $D = 438{,}784$ dimensions; the measurements say otherwise. The minimizer set

$$\Theta^\star \;=\; \big\{\theta \;:\; f_\theta|_{\mathcal{M}} = \alpha y + \beta \mathbf{1},\; \alpha \neq 0\big\} \tag{3}$$

imposes $m - 2$ conditions on D parameters whenever y is nonconstant on $\mathcal{M}$; wherever $\theta \mapsto f_\theta|_{\mathcal{M}}$ is a submersion—which we check rather than assume (App. E)—it is a submanifold of codimension exactly $m - 2$ (Prop. A.5). Over the evaluated images m runs from 82 to 18,966, so the codimension lies between 0.02% and 4.3% of D: $\Theta^\star$ retains at least 95% of the parameter dimensions on *every* evaluated image, even the densest KITTI frame. Adaptation drives $1 - \rho^2_{\mathcal{M}}$ from a median of 10^{-1} to 2×10^{-3}, yet dense accuracy keeps improving after the objective has effectively saturated (App. E): the run drifts along $\Theta^\star$, and that drift changes the unmeasured pixels while leaving the measured ones fitted.

This is the implicit-regularization regime of under-determined factorized regression (Gunasekar et al., 2017), with a generalization gap that is spatial rather than statistical. It makes a falsifiable prediction, confirmed in Sec. 5.4: interventions that improve the *optimization* or regularize in *parameter* space should not help, since no parameter-space penalty reduces the dimension of $\Theta^\star$. Shrinking $\Theta^\star$ requires information about the unmeasured pixels, and without ground truth there are two sources, the image and the frozen model's own prediction; SRP uses both, and only the second is defined far from every measurement.

4 Method

(C1) constrains our own design as much as the baseline's: any constant in data units—a penalty weight, an affinity bandwidth, a step length—can suit one benchmark only, so neither component below carries one. Fig. 2 shows where the two sit in one adaptation: SWS replaces the initialization, the adaptation loop is unchanged and SRP is a read-out applied to whichever prediction reaches it.

4.1 Semiparametric residual propagation

Let $\hat{f} = a f_\theta + b$ be the aligned prediction and $q_i = y_i - \hat{f}_i$ the residual at the measured pixels. The objective determines $\hat{f}$ on $\mathcal{M}$ and says nothing elsewhere, so we choose what happens elsewhere explicitly.

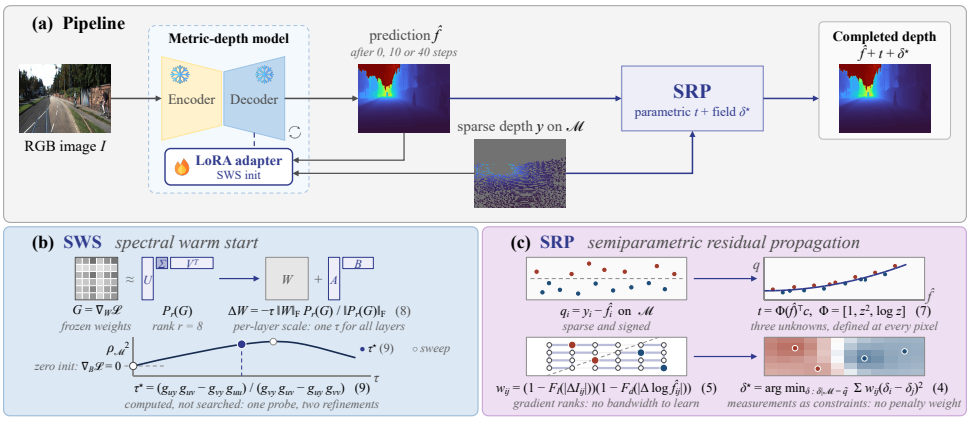

Figure 2: **Where SWS and SRP sit in one adaptation.** **(a)** Pipeline: a frozen monocular metric-depth model, a low-rank adapter initialized by SWS, and SRP as the read-out of whichever prediction reaches it; the loop runs 10 steps at the point we lead with, 40 at the baseline's budget and 0 when skipped. Images: the KITTI sample of Fig. 11, its sparse input and a 10-step run before and after the solve. **(b)** SWS: the step (8) and its amplitude (9), against a sweep of τ. **(c)** SRP: the residual on $\mathcal{M}$, the parametric term (7), the affinity (5) and the constrained field (4). Dashed outlines: frozen or inherited from Seo et al. (2026); color: ours.

A measurement is a constraint, not a preference. The obvious formulation trades a data term against a smoothness penalty, $\sum_{i\in\mathcal{M}}(\delta_i - q_i)^2 + \mu\sum_{(i,j)\in\mathcal{N}} w_{ij}(\delta_i - \delta_j)^2$, with $\mathcal{N}$ the four-neighborhood and w_{ij} an affinity that is large across a flat edge and small across a sharp one. Unlike the parameter-space problem, whose solution set has dimension close to D, this function-space problem has exactly one solution—the reason for moving the choice here. But it carries three constants in data units (μ, the bandwidth w_{ij} needs, and the solver's iteration count), which (C1) says will not transfer, and μ has no defensible value at all: a measurement is a value the field must take, not a term to trade against a prior. Imposing it as such replaces the data term by a constraint,

$$\delta^\star \;=\; \arg\min_{\delta\,:\,\delta|_{\mathcal{M}}=\tilde{q}} \;\sum_{(i,j)\in\mathcal{N}} w_{ij}\,(\delta_i-\delta_j)^2, \tag{4}$$

whose solution is graph-harmonic away from $\mathcal{M}$: the Dirichlet problem on the affinity graph with the measurements as boundary values ($\tilde{q}$ rather than q: one more term is subtracted below). **This is the same prior with its weight taken to the limit**—letting $\mu \to 0$ recovers (4) exactly (Prop. A.8)—and taking the limit improves MAE on the frozen prediction by 18% over a tuned μ.

An affinity with no bandwidth. The usual $w_{ij} = \exp(-\|I_i - I_j\|/\sigma)$ has the same defect one level down: σ is a threshold on image gradients, whose units the benchmark sets. We use each gradient's *rank* instead: with F_I and F_d the empirical distributions, over all edges of one image, of the image gradient and of the predicted log-depth gradient,

$$w_{ij} \;=\; \big(1 - F_I(|\Delta I_{ij}|)\big)\big(1 - F_d(|\Delta \log \hat{f}_{ij}|)\big), \tag{5}$$

so the flattest edge in the image has weight 1 and the sharpest 0, and an edge is crossable only if the image and the predicted geometry both say it is flat. Ranks are invariant to any monotone rescaling of the gradient—precisely what σ was there to absorb—so the construction has no constant at all, and it improves MAE by 12% over a tuned bandwidth.

The maximum principle is a ceiling, and it binds. A field harmonic on a graph is a convex combination of its boundary values: with h_{ij} the probability that a random walk started at i first meets $\mathcal{M}$ at j (Prop. A.8),

$$\delta_i \;=\; \sum_{j\in\mathcal{M}} h_{ij}\,\tilde{q}_j, \qquad h_{ij}\ge 0, \quad \textstyle\sum_j h_{ij} = 1, \tag{6}$$

so $\min_{\mathcal{M}} \tilde{q} \le \delta_i \le \max_{\mathcal{M}} \tilde{q}$ at every unmeasured pixel. The fact is classical (Levin et al., 2004; Zhu et al., 2003); we read it as a *ceiling* on what any propagated correction can do. **The solve can**

never correct by more than the measurements themselves show, and as i moves away from $\mathcal{M}$ the distribution $h_{i\cdot}$ spreads, so the correction flattens. Grouping pixels by distance to the nearest measurement shows where the ceiling binds (Tab. 12): indoors the solve stays within 0.97–$1.31\times$ of adaptation's squared error in every group, while DDAD is ahead within 4 pixels and $5.7\times$ worse between 16 and 64, and KITTI shows the same shape at less than half the size. The ceiling binds where the residual is both large and far from any measurement: the failure is one of reach, not of the prior.

A term the field cannot represent. What the field cannot supply is a correction defined far from every measurement, so we fit one on the measured pixels as a function not of position but of the predicted depth, which is available at every pixel:

$$t \;=\; \Phi(\hat{f})^{\top} c, \quad \Phi(z) = [\,1,\; z^2,\; \log z\,], \quad c = \arg\min_{c} \sum_{i \in \mathcal{M}} \big(\Phi(\hat{f}_i)^{\top} c - q_i\big)^2. \tag{7}$$

The affine alignment leaves q orthogonal on $\mathcal{M}$ to $\mathrm{span}\{1, \hat{f}\}$, so a fit on that span alone returns zero; $\hat{f}^2$ and $\log \hat{f}$ are the lowest-order terms not already exhausted, and the two the physics suggests: a monocular backbone's absolute error grows superlinearly with range, while its relative error is roughly constant. **This generalizes scale-and-shift alignment**, which assumes that the residual left by two degrees of freedom is unstructured; richer bases reach further but generalize worse, and the balance tips between three terms and four (App. F).

The solve then propagates only what the parametric term leaves: $\tilde{q}_i = q_i - t_i$ on $\mathcal{M}$ in (4), and the output is $\hat{f} + t + \delta^{\star}$. The constraint is untouched—at $i \in \mathcal{M}$ the output is $\hat{f}_i + t_i + (q_i - t_i) = y_i$—and the two pieces are complementary: one parametric, global and extrapolating, the other nonparametric, local and interpolating. What the term buys is reach: on DDAD it cuts the 16–64 pixel group from $5.7\times$ to $2.3\times$ and improves the nearer groups as well. **That it is a function of depth and not of position is essential**: the same three unknowns fitted on pixel coordinates—the regression-kriging design—leave that group almost untouched (App. F), because a trend on the grid says nothing about a far pixel that the solve does not. Indoors the term is nearly neutral, and on VOID, whose measured set is the most biased sample of its scene, it changes MAE by $+1.2\%$.

The iteration count is a budget, not a constant. In the penalized form the iteration count is a regularizer that must be tuned (App. H); for (4), accuracy is monotone in the work spent. We use a *cascadic* cycle, every level solving the true problem at its own resolution: a fixed 4×120 budget comes within 1.1% of the converged field, and its loop length depends on the resolution alone, so the solve is a static computation graph (Sec. 5.3).

4.2 Spectral warm start

SWS replaces the zero initialization with an analytic first step: *spectral* because its direction is a truncated singular value decomposition of the gradient, a *warm start* because it optimizes a surrogate rather than guessing a convenient size. Zero initialization suits fine-tuning on a corpus; with one image, a budget of tens of steps and (C3), it is a degenerate start. LoRA-GA also leaves that point but subtracts its initial product, so its run still starts at the frozen function; SWS takes the step.

Which step, in closed form. Let $G_l = \nabla_{W_l} \mathcal{L}$ be the gradient with respect to the *base* weight of an adapted layer at the frozen model. Among rank-r matrices of a given Frobenius norm, the one that decreases the loss fastest to first order maximizes $\langle -G_l, \Delta W \rangle$, and by von Neumann's trace inequality it is (Prop. A.6)

$$\Delta W_l \;=\; -\,\tau\, \|W_l\|_F \, \frac{\mathcal{P}_r(G_l)}{\|\mathcal{P}_r(G_l)\|_F}, \tag{8}$$

with $\mathcal{P}_r$ truncating to the leading r singular triplets. The factor $\|W_l\|_F$, absent from LoRA-GA, cannot be absorbed by any computed amplitude: τ is one scalar shared by *every* adapted layer, so per-layer scaling changes the direction of the joint step (App. G reports what dropping it costs).

How far, in closed form. The perturbation is linear in τ, so to first order the prediction on $\mathcal{M}$ along the ray is $u + \tau v$, with $u = \hat{f}|_{\mathcal{M}}$ and v the directional derivative, and (2) makes the objective

Table 1: **Five benchmarks, one protocol.** MAE and RMSE in meters; lower is better. "Steps" is the number of adaptation steps per image. **Bold and underlined** marks the best cell of the lower block; bold row labels mark SRP alone and the one we lead with. Rows that start from LoRA-TTO's initialization are means over five seeds; the others are deterministic. All rows come from one pipeline on one set of inputs (6,479 images, full splits), so none is comparable to a published figure.

		iBims-1		VOID		NYUv2		KITTI		DDAD	
Method	Steps	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Depth foundation models											
UniDepthV2 (Piccinelli et al., 2026)	0	0.302	0.395	0.238	0.594	0.306	0.429	1.539	3.560	5.176	14.843
+ scale & shift	0	0.100	0.216	0.184	0.555	0.145	0.253	1.405	2.977	5.774	13.866
Depth Anything V2 (Yang et al., 2024)	0	0.390	0.494	0.334	0.680	1.165	1.263	2.301	3.986	7.334	17.580
+ scale & shift	0	0.137	0.258	0.202	0.568	0.205	0.311	1.705	3.313	7.438	13.903
Depth Anything 3 (Lin et al., 2026)	0	1.112	1.222	0.737	1.049	0.979	1.081	2.200	3.772	16.004	24.504
+ scale & shift	0	0.110	0.222	0.190	0.555	0.158	0.264	1.462	3.041	4.631	10.476
Zero-shot depth completion, UniDepthV2 backbone											
Frozen decoder + scale & shift	0	0.100	0.216	0.191	0.574	0.173	0.280	1.419	3.017	5.888	14.088
+ TestPromptDC (Jeong et al., 2025)	—	0.053	0.165	0.170	0.534	0.134	0.236	1.123	2.367	5.973	13.212
+ LoRA-TTO (Seo et al., 2026)	40	0.041	0.146	0.136	0.505	0.054	0.125	0.438	1.372	2.002	5.937
+ SWS *(ours)*	40	0.040	0.147	0.133	0.502	0.052	0.124	0.399	1.357	1.752	5.846
+ **SRP** *(ours)*	**0**	0.040	0.162	0.152	0.539	0.055	0.142	0.340	1.448	1.480	6.325
+ **SWS** + **SRP** *(ours)*	**10**	0.041	**0.144**	0.134	0.504	0.054	0.125	0.321	**1.301**	1.386	**5.494**
+ LoRA-TTO + SRP *(ours)*	40	**0.037**	0.146	0.132	0.503	0.050	0.123	0.318	1.325	1.399	5.613
+ SWS + SRP *(ours)*	40	**0.037**	0.146	**0.130**	**0.501**	**0.049**	**0.122**	**0.310**	1.329	**1.341**	5.695

a function of $\rho_{\mathcal{M}}(u + \tau v, y)^2$ alone. With $\bar{\cdot}$ centering over $\mathcal{M}$ and $g_{pq} = \langle \bar{p}, \bar{q} \rangle$ the centered inner product of any two of u, v and y, that correlation is a ratio of quadratics in τ whose unique maximizer is (Prop. A.7)

$$\tau^\star = \frac{g_{uy}\, g_{uv} - g_{vy}\, g_{uu}}{g_{vy}\, g_{uv} - g_{uy}\, g_{vv}}. \tag{9}$$

One extra forward pass supplies v by finite difference, and the probe size cancels to first order. A tangent taken at the origin undershoots when the ray curves (0.84 of the true maximizer indoors and 0.55 on DDAD, over 60 images), so we re-solve the same model from where it lands, a secant through the last two points evaluated. Two refinements beat none and three, for 4 forward passes in total; this is the only integer in the method chosen on accuracy, fixed once for all five benchmarks, and it beats every greedy amplitude search we tried (App. G). The adapter realizes (8) as sAB from the truncated singular vectors; both factors are then free and the shipped recipe runs unchanged, so a warm-started run differs from the reference only in its initialization.

5 Experiments

5.1 Protocol

Five benchmarks differ in scene type, depth range and—most importantly here—how the sparse measurements are produced: **NYUv2** (Silberman et al., 2012) and **iBims-1** (Koch et al., 2018) (indoor, structured light and laser scan), **KITTI** (Uhrig et al., 2017) and **DDAD** (Guizilini et al., 2020) (outdoor driving, lidar), and **VOID** (Wong et al., 2020) (visual-inertial odometry points). Sparse inputs follow the protocol of Seo et al. (2026); the measured set runs from 82 to $18{,}966$ pixels, a factor of 231. We evaluate every image of every split at native resolution, 6,479 in total, after removing the 25 images that an earlier version of the amplitude rule used for calibration (App. C). The splits are unequal—DDAD alone is 69% of the four benchmarks other than VOID, which we call the core—so the headline is the unweighted mean of per-benchmark means, and we name the pooled per-image statistic wherever a paired test needs it.

All rows come from one pipeline on one set of inputs: TestPromptDC (Jeong et al., 2025) is re-implemented and run under our protocol, while the diffusion-prior methods (Viola et al., 2025; Lee et al., 2025) are not run, as sampling a diffusion model on every image is beyond our compute budget. The frozen backbone is UniDepthV2 with a ViT-L encoder (Piccinelli et al., 2026), with rank-8 adapters in the decoder's linear layers and the configuration published with LoRA-TTO: 40 steps of AdamW (Loshchilov & Hutter, 2019) at a learning rate of 10^{-2}. Our runs use the same optimizer, learning rate and, at the full budget, step count, differing only in the initialization and the read-out. The forward pass is bitwise reproducible and backward reductions are deterministic; rows that start from LoRA-TTO's initialization are means over five seeds, and SWS is deterministic.

5.2 MAIN RESULT

Tab. 1's upper block reports three metric-depth backbones raw and after a scale-and-shift fit to the same measurements; everything below is measured on top of that alignment. **The lower block is a cost curve, not a ranking**: rows differ in what they spend as well as in what they achieve. Relative differences are signed, so negative is an improvement.

At equal budget, the read-out improves every benchmark on both metrics. Adding SRP to the shipped 40-step run improves both metrics on all five benchmarks, and MAE on 100.0% of the core's images. The mean of per-benchmark relative MAE is -18.5% over the core and -15.3% over all five; per image, which is what a paired bootstrap applies to, it is -26.5% ($[-26.7, -26.2]$, $p < 10^{-4}$). The two differ because DDAD carries both the most images and the largest gain (App. D). The smallest gains are on RMSE for VOID and iBims-1, below half a percent (the table rounds the latter to a tie): a squared metric prices the tail, which the solve reaches least.

At a quarter of the budget, it matches or beats the baseline everywhere. SWS with 10 steps, read out through SRP, is the configuration we lead with: at least as accurate as the full 40-step LoRA-TTO on all five benchmarks and both metrics, at a quarter of the steps and -15.1% MAE over the core. Indoors it ties or wins narrowly—no cell moves by as much as 2%, and on 4 of the 10 cells the paired per-image difference is not distinguishable from zero ($p \geq 0.08$); on the two driving benchmarks it lowers MAE by 27% and 31% and RMSE by 5% and 7%, with every remaining cell at $p \leq 0.0038$. **Where the gain falls is predicted**: KITTI and DDAD have the largest scenes and the largest remaining errors, where the choice among minimizers decides most (Sec. 3.2) and the maximum principle binds (Tab. 12). Nor is it a density effect alone: its Spearman correlation is $+0.34$ with a sample's measurement density but -0.50 with the baseline's error on it (App. D).

With no adaptation at all, the read-out is a different operating point. Applied to the frozen prediction, SRP needs no backward pass, adapter or optimizer, and is ahead of LoRA-TTO on MAE on three of the five benchmarks. It is behind on RMSE on all five, because it leaves a tail it cannot reach and a squared metric is priced by that tail.

The components transfer across backbones. Because Sec. 3 concerns the objective rather than the backbone, App. D repeats the leading rows on MoGe-2 with nothing retuned: the 40-step SWS + SRP result carries to both backbones and wins every driving cell on each, and at 10 steps the driving gains persist while a few indoor cells move the wrong way by at most $+1.1\%$. **The accuracy/cost curve transfers; where a given backbone sits on it does not.**

5.3 COST

Every method is timed by us on one NVIDIA RTX 4090 under one protocol (Tab. 7). **The adaptation was the cost, and it no longer is.** Everything SRP adds—affinity, fit and captured solve—totals 21 ms at VGA, against 88 ms for the unavoidable forward pass and 4,343 ms for a full 40-step adaptation, and stays at 5–24% of the forward pass across resolutions. The remaining cost is therefore the backbone's: SRP on the frozen prediction is 30–40× faster end to end than the adaptation, and the 10-step configuration 2.5–2.7× faster at equal or better accuracy. Because (4)'s loop length is fixed by the resolution alone, we capture the solve in a CUDA graph and replay it, reproducing the uncaptured result bit for bit at 5–10× lower cost (App. D).

Accuracy and cost on one pair of axes. In Fig. 3 the three points we ship form a *frontier*—no method we compare against lies below it on either metric—whose shape differs between metrics: on MAE it falls monotonically with budget from the adaptation-free point, while on RMSE the adaptation-free point sits *above* the baseline and the 10-step point is the first below it. **That crossing is why we report a budget rather than an operating point**: which point to deploy depends on the latency available and on how heavily the tail is priced, while the baseline's single point is dominated by the 10-step configuration.

5.4 WHAT EACH PART IS WORTH, AND WHAT DOES NOT WORK

Each component earns its place at a different end of the curve. Removing one design decision from SRP at a time (App. F): constraints, ranked affinities and image structure are worth 18%, 12% and 27% on the frozen prediction, and all three are within 1.6% of full SRP after 40 steps: they

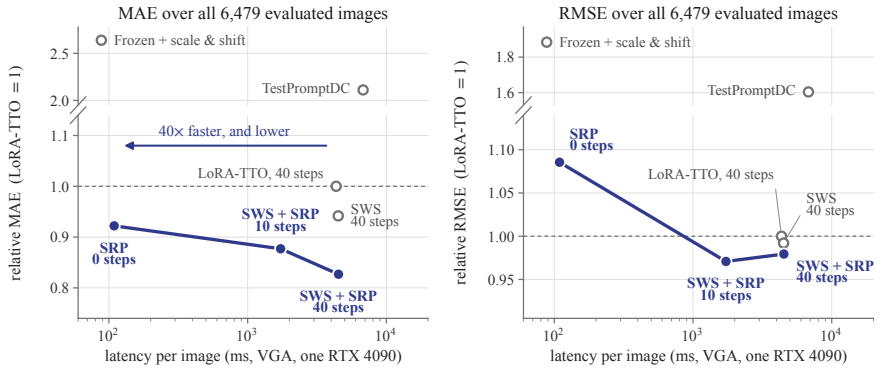

Figure 3: **Tab. 1 and Tab. 7 on one pair of axes.** Each point is a configuration's error relative to LoRA-TTO (mean over the five benchmarks, all 6,479 images) against its latency at VGA on one RTX 4090; the dashed line is the 40-step baseline, and lower-left is better. **The y axis is broken** between the adapted and unadapted configurations.

carry information a long way, and after adaptation the residual is small and local. The parametric term is the sharpest case: it is worth 13% of RMSE on DDAD with no adaptation, and after 40 steps it changes no benchmark's MAE in the third decimal place. SWS is the mirror image (App. G): aligning the initialization with the gradient's leading singular subspace, as LoRA-GA does, is worth 3.9% at 40 steps and 8.6% at 10—the steps a zero-initialized run spends leaving the degenerate point of (C3) are a fixed cost—and taking the step adds 3.4% and 5.3%. Every SWS variant has a seed standard deviation of exactly zero, against 0.049 for the baseline at 10 steps.

The analysis makes a falsifiable prediction, and it holds. Sec. 3.2 predicts that interventions improving the optimization, or regularizing in parameter space, should not help. We ran twelve, each applied to LoRA-TTO with everything else fixed—schedules, factor-asymmetric learning rates, sharpness-aware minimization, averaging, early stopping, proximal and trust-region terms among them. Ten come in at or above parity; the only two below it move along the exactly flat directions of (C2), and even they are within half a percent, as expected for changes the objective cannot see. Two further negatives sharpen the method. Restricting the adapter to the leading singular subspace of G_l reaches 1.15× the baseline's MAE, so SWS frees both factors after the spectral step. And no training loss folds the solve into the adapter: asked only to *copy* the converged field, the adapter reaches 1.737 against the field's 1.123 (relative to SWS + SRP), so **the minimizer that (4) selects lies outside what a rank-8 adapter can represent**, and the two stages stay separate (App. H).

6 Conclusion

The test-time objective of low-rank depth completion is exactly a correlation criterion scaled by the variance of the measurements. Its gradients therefore carry the units of the data, its standard initialization is degenerate and its minimizers retain at least 95% of the parameter dimensions on every image, so the difficulty is choosing a minimizer, not reaching one. SWS computes the first step, and SRP chooses among the minimizers in function space, with the measurements as constraints, an affinity with no length scale and a parametric term in the predicted depth. Together they yield an accuracy/cost curve—a read-out with no backward pass, a 10-step configuration that matches or beats the 40-step baseline on every benchmark and metric, and an equal-budget configuration that improves on it everywhere—that transfers across frozen backbones. The analysis was tested rather than asserted: optimization and parameter-space interventions do not help, as it predicts, and the maximum principle binds exactly where residuals are large and far from any measurement.

Limitations. Without adaptation, SRP still trails the 40-step run 4 to 16 pixels from a measurement (Tab. 12), and VOID's tracker-placed points are a biased sample of the scene (App. J). Affinities robust to such feature-driven sparsity and an extrapolating field that converges at this budget are natural next steps, both about how to choose the unmeasured pixels—the question the closed form identifies as the one that matters.

AI USE STATEMENT

We used AI tools to assist with language editing, including grammar correction and refinement of the academic writing style, as well as ablation experiment design and ablation code implementation. The core research ideas, methodology, and main experiments were developed and conducted by the authors. All AI-assisted outputs were reviewed and tested by the authors, who take full responsibility for the final manuscript and code.

REPRODUCIBILITY STATEMENT

Analysis. Every statement of Sec. 3 and Sec. 4 is proved in App. A, and App. E checks on every sample of a diagnostic subset that the implemented objective is the one analyzed, including the envelope identity, both flat manifolds of (C2) over the full gauge group, the vanishing curvature block of (C3) and the submersion hypothesis behind Sec. 3.2. **Method.** App. B gives the linear system of (4), the cascadic hierarchy, the parametric fit and pseudocode for SWS (Alg. 1); Tab. 3 lists the full configuration, which has no calibration set and no constant in data units, and whose two integers (the cascadic budget and the refinement count) are fixed by the measurements in Sec. 4.1 and App. G. The captured and uncaptured solves agree bitwise, the forward pass is bitwise reproducible and backward reductions are deterministic. **Evaluation.** App. C states the datasets, the per-benchmark image counts, the removal of the 25 calibration images, the metrics and statistics and the timing protocol. Tab. 1 and Tab. 4 evaluate every image of every split (6,479 images), all rows come from one pipeline on one set of inputs and every latency is measured by us; the other studies run on stratified subsets of the splits. Negative results are reported with their sample sizes and sweep extents in App. H. **Code.** Source code for the method, the benchmark harness and every table and figure generator is included in the supplementary material, with a provenance file mapping each table and figure to the script and result files that produce it; three result sets (the foundation-model and TestPromptDC rows of Tab. 1, the TestPromptDC row of Tab. 7 and Tab. 20) regenerate from shipped intermediates rather than raw inputs.

REFERENCES

Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *ECCV*, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *CVPR*, 2020.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *NeurIPS*, 2017.

Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM Monographs on Mathematical Modeling and Computation. SIAM, 1998.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *ICML*, 2024.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *ECCV*, 2010.

Tomislav Hengl, Gerard B. M. Heuvelink, and David G. Rossiter. About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301–1315, 2007.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Chanhwi Jeong, Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Test-time prompt tuning for zero-shot depth completion. In *ICCV*, 2025.

Bingxin Ke, Qunjie Zhou, Jiahui Huang, Xuanchi Ren, Tianchang Shen, Konrad Schindler, Laura Leal-Taixé, and Shengyu Huang. Depth completion as parameter-efficient test-time adaptation. *arXiv preprint arXiv:2602.14751*, 2026.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *ECCV Workshops*, 2018.

Hyoseok Lee, Kyeong Seon Kim, Byung-Ki Kwon, and Tae-Hyun Oh. Zero-shot depth completion via test-time alignment with affine-invariant depth prior. In *AAAI*, 2025.

Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. *ACM TOG*, 23(3): 689–694, 2004.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *ICML*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *NeurIPS*, 2024.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *ECCV*, 2020.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE TPAMI*, 48(3):2354–2367, 2026.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2022.

Minseok Seo, Wonjun Lee, Jaehyuk Jang, and Changick Kim. Efficient test-time optimization for depth completion via low-rank decoder adaptation. *arXiv preprint arXiv:2603.01765v4*, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *3DV*, 2017.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-DC: Zero-shot monocular depth completion with guided diffusion. In *ICCV*, 2025.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *NeurIPS*, 2025.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *NeurIPS*, 2024.

Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE RA-L*, 5(2):1899–1906, 2020.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *NeurIPS*, 2024.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *ICML*, 2024.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.

APPENDIX

The appendix is organized as follows.

- **App. A** proves the statements of Sec. 3 and Sec. 4.
- **App. B** gives the implementation of SRP and SWS, with pseudocode.
- **App. C** describes datasets, splits, metrics, hyperparameters, baselines and the timing protocol.
- **App. D** extends the main results: all four metrics, paired statistics, a second backbone, the budget sweep and the latency breakdown.
- **App. E** verifies the analysis numerically and measures the size of its three consequences.
- **App. F** and **App. G** ablate SRP and SWS, one design decision at a time.
- **App. H** reports design alternatives that do not help, first those the analysis predicts should fail and then others a reader may consider.
- **App. I** shows qualitative results, and **App. J** discusses limitations.

Diagnostic studies run on stratified subsets of the evaluation splits. They are read down a column rather than against Tab. 1, whose rows are the full splits.

A DERIVATIONS

This section proves the statements of Sec. 3 and Sec. 4. For $p \in \mathbb{R}^m$ indexed by $\mathcal{M}$, let $\bar{p} = p - \frac{1}{m}(\mathbf{1}^\top p)\mathbf{1}$ be its centered version, $\mathrm{Var}_{\mathcal{M}}(p) = \frac{1}{m}\|\bar{p}\|^2$, $\mathrm{Cov}_{\mathcal{M}}(p,q) = \frac{1}{m}\langle \bar{p}, \bar{q}\rangle$ and $\rho_{\mathcal{M}}(p,q) = \mathrm{Cov}_{\mathcal{M}}(p,q)/\sqrt{\mathrm{Var}_{\mathcal{M}}(p)\mathrm{Var}_{\mathcal{M}}(q)}$. For fixed θ we write $f = f_\theta|_{\mathcal{M}}$.

Proposition A.1 (Closed form, Eq. (2))**.** *If $\mathrm{Var}_{\mathcal{M}}(f) > 0$, the inner problem of (1) has the unique minimizer $a^\star = \mathrm{Cov}_{\mathcal{M}}(f,y)/\mathrm{Var}_{\mathcal{M}}(f)$, $b^\star = \frac{1}{m}\mathbf{1}^\top(y - a^\star f)$, and $\mathcal{L}(\theta) = \mathrm{Var}_{\mathcal{M}}(y)\left(1 - \rho_{\mathcal{M}}(f,y)^2\right)$.*

Proof. For fixed a the optimal b is the mean of $y - af$, which leaves $\frac{1}{m}\|a\bar{f} - \bar{y}\|^2 = a^2\mathrm{Var}_{\mathcal{M}}(f) - 2a\mathrm{Cov}_{\mathcal{M}}(f,y) + \mathrm{Var}_{\mathcal{M}}(y)$. This is strictly convex in a when $\mathrm{Var}_{\mathcal{M}}(f) > 0$ and is minimized at $a^\star$, with value $\mathrm{Var}_{\mathcal{M}}(y) - \mathrm{Cov}_{\mathcal{M}}(f,y)^2/\mathrm{Var}_{\mathcal{M}}(f) = \mathrm{Var}_{\mathcal{M}}(y)\,(1 - \rho^2_{\mathcal{M}})$. □

Proposition A.2 (The detached gradient is exact)**.** *Let $\ell(\theta,a,b) = \frac{1}{m}\sum_{i\in\mathcal{M}}(a[f_\theta]_i + b - y_i)^2$ with f_θ differentiable in θ. On $\{\theta : \mathrm{Var}_{\mathcal{M}}(f_\theta) > 0\}$, $\nabla\mathcal{L}(\theta) = \nabla_\theta\,\ell(\theta,a,b)\big|_{(a,b)=(a^\star(\theta),\,b^\star(\theta))}$.*

Proof. By Prop. A.1, $a^\star$ and $b^\star$ are differentiable functions of θ on this set. The chain rule gives $\nabla\mathcal{L} = \nabla_\theta\ell + \partial_a\ell\,\nabla a^\star + \partial_b\ell\,\nabla b^\star$, and $\partial_a\ell = \partial_b\ell = 0$ at the inner minimizer. □

Proposition A.3 (Invariances, (C2))**.** *(i) $\mathcal{L}$ is unchanged when f_θ is replaced by $\alpha f_\theta + \beta\mathbf{1}$ for any $\alpha \neq 0$ and β. (ii) For an adapted layer with $\Delta W = sAB$, $A \in \mathbb{R}^{d_{\mathrm{out}}\times r}$ and $B \in \mathbb{R}^{r\times d_{\mathrm{in}}}$, the action $(A,B) \mapsto (AM, M^{-1}B)$ of $M \in \mathrm{GL}(r)$ leaves ΔW, hence $\mathcal{L}$, unchanged, and the orbit of any (A,B) with $\mathrm{rank}\,A = r$ is an r^2-dimensional submanifold.*

Proof. (i) Centering removes β, and α multiplies $\mathrm{Cov}_{\mathcal{M}}$ by α and $\sqrt{\mathrm{Var}_{\mathcal{M}}}$ by $|\alpha|$, so $\rho^2_{\mathcal{M}}$ is unchanged. (ii) $AMM^{-1}B = AB$. If $\mathrm{rank}\,A = r$, then $AM = A$ forces $M = I$, so the stabilizer is trivial and the orbit has the dimension of $\mathrm{GL}(r)$. □

Proposition A.4 (The zero initialization, (C3))**.** *Let G be the gradient of $\mathcal{L}$ with respect to the base weight of an adapted layer at the frozen model, and consider any point with $A = 0$. Then (i) $\mathcal{L}$ is constant in B on $\{A = 0\}$, so $\nabla_B\mathcal{L} = 0$ and $\nabla^2_{BB}\mathcal{L} = 0$ there; (ii) $\nabla_A\mathcal{L} = s\,GB^\top$; and (iii) if $G \neq 0$, the Hessian of $\mathcal{L}$ with respect to (A,B) is indefinite.*

Proof. (i) $\Delta W = sAB = 0$ for every B when $A = 0$. (ii) This is the chain rule through $\Delta W = sAB$, with $\nabla_{\Delta W}\mathcal{L} = G$ at $\Delta W = 0$. (iii) Perturb $(0,B)$ to $(\varepsilon a, B + \varepsilon b)$. Then $\Delta W = s\varepsilon\,aB + s\varepsilon^2 ab$, and the second-order term of $\mathcal{L}$ is $\varepsilon^2 Q(a,b)$ with $Q(a,b) = s\langle G, ab\rangle + \frac{s^2}{2}\,\mathcal{H}[aB, aB]$, where

$\mathcal{H}$ is the Hessian of $\mathcal{L}$ in W. Choose $a = ue_1^\top$ and $b = e_1 v^\top$ with $u^\top G v \neq 0$, so that $\langle G, ab\rangle \neq 0$. As $t \to 0$, $Q(ta, \pm b) = \pm\, t\, s\langle G, ab\rangle + O(t^2)$ takes both signs. □

Proposition A.5 (Dimension of the minimizer set)**.** *Let $F : \mathbb{R}^D \to \mathbb{R}^m$, $\theta \mapsto f_\theta|_{\mathcal{M}}$, be smooth, let y be nonconstant on $\mathcal{M}$, and let $\Theta^\star = F^{-1}(S)$ with $S = \{\alpha y + \beta \mathbf{1} : \alpha \neq 0,\ \beta \in \mathbb{R}\}$. Then $\Theta^\star$ is exactly the zero set of $\mathcal{L}$, and if it is nonempty and F is a submersion at each of its points, it is an embedded submanifold of dimension $D - m + 2$.*

Proof. $\mathcal{L} = 0$ if and only if $\rho^2_{\mathcal{M}} = 1$, i.e., $\bar{f} \parallel \bar{y}$, i.e., $f \in S$ ($\alpha \neq 0$ because $\mathrm{Var}_{\mathcal{M}}(f) > 0$). Since y and $\mathbf{1}$ are linearly independent, S is an embedded two-dimensional submanifold of $\mathbb{R}^m$. A submersion is transverse to every submanifold, so $F^{-1}(S)$ is an embedded submanifold of codimension $m - 2$. □

With $D = 438{,}784$ and $m \in [82, 18{,}966]$, the codimension $(m-2)/D$ lies between 0.02% and 4.3%, which is the statement of Sec. 3.2; App. E tests the submersion hypothesis.

Proposition A.6 (Steepest rank-r step, Eq. (8))**.** *For $G \neq 0$ and $c > 0$, $\max\{\langle -G, X\rangle : \mathrm{rank}\, X \leq r,\ \|X\|_F = c\} = c\,\|\mathcal{P}_r(G)\|_F$, attained at $X = -c\,\mathcal{P}_r(G)/\|\mathcal{P}_r(G)\|_F$.*

Proof. By von Neumann's trace inequality and the Cauchy–Schwarz inequality over the at most r nonzero singular values of X, $\langle -G, X\rangle \leq \sum_{i \leq r} \sigma_i(G)\,\sigma_i(X) \leq \left(\sum_{i \leq r} \sigma_i(G)^2\right)^{1/2} \|X\|_F = c\,\|\mathcal{P}_r(G)\|_F$, and $\langle G, \mathcal{P}_r(G)\rangle = \|\mathcal{P}_r(G)\|_F^2$ shows that the bound is attained. □

Eq. (8) is this maximizer with $c = \tau \|W_l\|_F$.

Proposition A.7 (The amplitude, Eq. (9))**.** *Let $u, v, y \in \mathbb{R}^m$ with $\bar{u}$ and $\bar{v}$ linearly independent, $g_{pq} = \langle \bar{p}, \bar{q}\rangle$, and $R(\tau) = \rho_{\mathcal{M}}(u + \tau v, y)^2$. If $\Delta = g_{vy} g_{uv} - g_{uy} g_{vv} \neq 0$, then $\tau^\star = (g_{uy} g_{uv} - g_{vy} g_{uu})/\Delta$ is the unique global maximizer of R on $\mathbb{R}$. If $\Delta = 0$, R has no maximizer on $\mathbb{R}$.*

Proof. $R(\tau)$ is the squared cosine between $\bar{u} + \tau\bar{v}$ and $\bar{y}$. Over the plane $\Pi = \mathrm{span}\{\bar{u}, \bar{v}\}$, the squared cosine with $\bar{y}$ is maximized exactly on the nonzero multiples of the projection $p = \alpha\bar{u} + \beta\bar{v}$ of $\bar{y}$ onto Π, where Cramer's rule, with $\det = g_{uu} g_{vv} - g_{uv}^2 > 0$, gives $\alpha \det = g_{vv} g_{uy} - g_{uv} g_{vy} = -\Delta$ and $\beta \det = g_{uu} g_{vy} - g_{uv} g_{uy}$. Every vector of Π with nonzero $\bar{u}$-coefficient is a multiple of $\bar{u} + \tau\bar{v}$ for exactly one τ, and squared cosines are invariant to scaling. If $\Delta \neq 0$, then $\alpha \neq 0$, so p is reached at $\tau^\star = \beta/\alpha$, which is the stated ratio, and at no other τ. If $\Delta = 0$, then $p \parallel \bar{v}$ is approached only as $|\tau| \to \infty$. □

The case $\Delta = 0$ is the guard of App. B.

Proposition A.8 (The constrained solve, Eqs. (4) and (6))**.** *Let $\mathbf{L} = \mathbf{D} - \mathbf{W}$ be the Laplacian of the affinity graph, U the unmeasured pixels and $P = \mathbf{D}^{-1}\mathbf{W}$, and suppose every pixel of U has a positive-weight path to $\mathcal{M}$. Then (i) $\mathbf{L}_{UU}$ is symmetric positive definite and (4) has the unique solution $\delta_U = \mathbf{L}_{UU}^{-1}\mathbf{W}_{U\mathcal{M}}\,\tilde{q}$; (ii) $\delta_U = H\tilde{q}$ with $H = (I - P_{UU})^{-1} P_{U\mathcal{M}}$ entrywise nonnegative and $H\mathbf{1} = \mathbf{1}$, where H_{ij}, the h_{ij} of (6), is the probability that the random walk with transition matrix P started at i first enters $\mathcal{M}$ at j; and (iii) the minimizer of the penalized objective $\sum_{i \in \mathcal{M}} (\delta_i - \tilde{q}_i)^2 + \mu\,\delta^\top \mathbf{L}\delta$ converges to $\delta^\star$ as $\mu \to 0$.*

Proof. (i) The stationarity condition of (4) in δ_U is $\mathbf{L}_{UU}\delta_U = -\mathbf{L}_{U\mathcal{M}}\tilde{q} = \mathbf{W}_{U\mathcal{M}}\tilde{q}$. $\mathbf{L}_{UU}$ is symmetric and weakly diagonally dominant with a nonpositive off-diagonal, and every connected component of U contains a pixel adjacent to $\mathcal{M}$, whose row is strictly dominant; hence $\mathbf{L}_{UU}$ is nonsingular, and, being symmetric with a nonnegative spectrum by Gershgorin's theorem, positive definite. (ii) $\mathbf{L}_{UU} = \mathbf{D}_{UU}(I - P_{UU})$, and P_{UU} is sub-stochastic with spectral radius below one under the connectivity assumption, so $(I - P_{UU})^{-1} = \sum_{k \geq 0} P_{UU}^k \geq 0$; this series is the first-passage decomposition of the walk, and $P_{UU}\mathbf{1} + P_{U\mathcal{M}}\mathbf{1} = \mathbf{1}$ gives $H\mathbf{1} = \mathbf{1}$. (iii) The penalized minimizer solves $(\mathbf{L} + \mu^{-1}\Pi_{\mathcal{M}})\,\delta = \mu^{-1}\Pi_{\mathcal{M}}\tilde{q}$, with $\Pi_{\mathcal{M}}$ the coordinate projection onto $\mathcal{M}$ and $\tilde{q}$ extended by zero; as $\mu \to 0$ its restriction to $\mathcal{M}$ tends to $\tilde{q}$ and its restriction to U tends to the solution of (i). □

Orthogonality of the aligned residual. The normal equations of the inner problem give $\sum_{i\in\mathcal{M}} q_i = 0$ and $\sum_{i\in\mathcal{M}} q_i \hat{f}_i = 0$, so a least-squares fit of q on $\mathrm{span}\{1, \hat{f}\}$ alone returns zero. This is why (7) starts from $\hat{f}^2$ and $\log \hat{f}$; adding $\hat{f}$ to them does change the joint fit, and App. F measures that variant.

B Method Details

B.1 Semiparametric residual propagation

The linear system. Problem (4) is assembled on the full-resolution grid over the four-neighbor graph, with $\mathbf{W}$ the matrix of affinities (5), $\mathbf{D}$ its row sums and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ (we avoid L, D and W, which denote the loss, the adapter dimension and a frozen weight matrix). The measured pixels are eliminated rather than penalized, so the solved system is $\mathbf{L}_{UU}\,\delta_U = -(\mathbf{L}\,\tilde{q})|_U$ with $\delta_{\mathcal{M}} = \tilde{q}$ substituted in. By Prop. A.8 it is symmetric positive definite exactly when every pixel of U has a positive-weight path to $\mathcal{M}$. The affinity (5) assigns weight zero only to the globally sharpest edge of each of its two factors, so at most two of the roughly $2N$ edges are cut; two cut edges can disconnect a four-neighbor grid only by isolating a corner pixel incident to both, a coincidence of probability zero for continuous gradients. We solve the system with conjugate gradients preconditioned by the diagonal of $\mathbf{L}_{UU}$.

The cascadic hierarchy. Coarse levels are built by max-pooling the constraint mask—a coarse cell is measured if any of its fine cells is, and takes the mean of those measurements—and by area-averaging the edge weights. Every level runs the same fixed number of conjugate-gradient iterations on its own grid and is prolonged bilinearly to the next, so coarse levels supply a starting point and never an operator the fine level has to trust. The deployed budget is 4 levels of 120 iterations.

The parametric term. The fit (7) is a 3×3 normal-equation solve in double precision, fitted on the measured pixels before the field is solved and added back afterward. It is evaluated at $\hat{f}$ clamped to the range $\hat{f}$ takes on $\mathcal{M}$, so the $\hat{f}^2$ and $\log \hat{f}$ terms never extrapolate past depths the fit has seen.

Graph capture. Because the number of iterations at every level is fixed by the resolution alone, the whole solve is a static computation graph: we capture it once per input shape as a CUDA graph and replay it, which removes the per-iteration kernel-launch overhead that otherwise dominates its cost (App. D). On every sample of the calibration set, we verify that the captured and uncaptured solutions agree bitwise.

B.2 Spectral warm start

Alg. 1 summarizes SWS. The gradient G_l of every adapted layer comes from one backward pass at the frozen weights, and the amplitude from forward passes along the ray, all charged once per image; the deployed configuration uses 4 forward passes in total (Tab. 16).

Algorithm 1 Spectral warm start (SWS) for one image.

Require: frozen model with adapted layers $\{W_l\}$, measurements y on $\mathcal{M}$, rank r, number of refinements n (deployed: $n = 2$)

$G_l \leftarrow \nabla_{W_l}\mathcal{L}$ for every adapted layer ▷ one backward pass

$V_l \leftarrow -\|W_l\|_F\,\mathcal{P}_r(G_l)/\|\mathcal{P}_r(G_l)\|_F$ ▷ direction of (8)

$u \leftarrow \hat{f}|_{\mathcal{M}}$ at $\tau = 0$; $v \leftarrow$ finite-difference derivative of $\hat{f}|_{\mathcal{M}}$ along $\{V_l\}$ ▷ one probe

$\tau \leftarrow \tau^\star(u, v, y)$ by (9), or 0 if its denominator is zero

for $k = 1, \ldots, n$ **do**

  evaluate $\hat{f}|_{\mathcal{M}}$ at τ; $(u, v) \leftarrow$ value at τ and secant slope through the last two evaluated points

  $\tau \leftarrow \tau + \tau^\star(u, v, y)$ ▷ same model, re-solved where it lands

end for

factor τV_l into sAB from its truncated singular vectors; free both factors

return the initialized adapter; run the unchanged T-step adaptation, then SRP

One guard on (9). The denominator of (9) vanishes exactly when the first-order model has no maximizer along the ray (Prop. A.7). The implementation then takes $\tau = 0$, the frozen initialization, rather than the unbounded step the tangent would license, so a degenerate ray costs the run its warm start and nothing else. It is the only branch in SWS and not a tuned threshold: the test is exact equality to zero in single precision.

C Experimental Setup

Datasets and sparse inputs. Tab. 2 reports the measured set $\mathcal{M}$ per benchmark. The sparse inputs follow the protocol of Seo et al. (2026) and differ in kind rather than only in density: NYUv2 and iBims-1 are uniform samples of a dense sensor, KITTI and DDAD are raw lidar returns projected into the image, and VOID's points are whatever its visual-inertial front end tracks. This is why VOID alone spans more than an order of magnitude in m within one benchmark, and why its measured set concentrates on corners and textured structures (App. J). Across all 6,479 images m runs from 82 to 18,966, a factor of 231.

Table 2: **The measured set $\mathcal{M}$ per benchmark.** Over all 6,479 evaluated images, m is the number of measured pixels and the share is m over the image's pixel count.

Benchmark	Images	measured pixels m		share of the image	
		median	range	median	range
iBims-1	95	974	961–993	0.32%	0.32–0.33%
VOID	795	702	82–1,373	0.23%	0.03–0.46%
NYUv2	649	475	455–488	0.72%	0.69–0.74%
KITTI	995	16,675	12,030–18,966	3.96%	2.85–4.50%
DDAD	3,945	4,782	1,989–5,788	0.47%	0.20–0.57%

Splits and the calibration set. The 6,479 evaluated images are 95 iBims-1, 795 VOID, 649 NYUv2, 995 KITTI and 3,945 DDAD, every image of every split at native resolution. An earlier version of the amplitude rule, reported as an ablation in Tab. 16, placed its grid on a calibration set drawn from the same five sources; several diagnostics also use it. It was drawn to exclude the evaluation keys of a smaller evaluation set, so once every image of every split is evaluated the two overlap, and we remove the 25 calibration images from the evaluated set before any number is computed (6,479 of 6,504). The final configuration has nothing to calibrate, and no selection step reads the dense ground truth of a reported sample.

Metrics and statistics. MAE (mean absolute error) and RMSE (root-mean-square error) are $\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}|\hat{d}_i - y_i^{\mathrm{gt}}|$ and $\left(\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}(\hat{d}_i - y_i^{\mathrm{gt}})^2\right)^{1/2}$ in meters, with $\hat{d}$ the dense output and $\mathcal{G}$ the pixels with valid ground truth; AbsRel (absolute relative error) is the mean of $|\hat{d}_i - y_i^{\mathrm{gt}}|/y_i^{\mathrm{gt}}$, and δ_1 the fraction of pixels with $\max(\hat{d}_i/y_i^{\mathrm{gt}}, y_i^{\mathrm{gt}}/\hat{d}_i) < 1.25$. Every benchmark mean weights its images equally. Relative figures are signed so that negative is an improvement; aggregates are unweighted means of per-benchmark relatives unless stated as pooled per image, in which case confidence intervals and p-values come from a paired bootstrap over images.

Baselines. The foundation-model rows of Tab. 1 run each backbone's released model on the same inputs, raw and after a least-squares scale-and-shift fit to the measurements. TestPromptDC (Jeong et al., 2025) is re-implemented and run under our protocol. LoRA-TTO (Seo et al., 2026) uses its released configuration (Tab. 3) and is averaged over five seeds. The diffusion-prior methods (Viola et al., 2025; Lee et al., 2025) are not run, as sampling a diffusion model on every evaluated image is beyond our compute budget.

Timing protocol. Every latency is measured by us on one NVIDIA RTX 4090, each configuration in its own process: 10 warm-up images, then 30 timed images, reported as the median, with regions bracketed by `cuda.synchronize`. The nominal resolutions VGA, SD, HD and FHD are cropped to a multiple of the backbone's patch size, giving 476×630, 476×714, 714×1274 and 1078×1918.

Table 3: **Configuration.** Everything not listed is inherited unchanged from the released LoRA-TTO configuration (Seo et al., 2026); neither component adds a constant in data units.

Component	Setting	Value
Backbone	frozen model	UniDepthV2 with a ViT-L encoder (Piccinelli et al., 2026)
	second backbone	MoGe-2 (Wang et al., 2025), in Tab. 5
Adapter	placement, rank	decoder linear layers (MoGe-2: neck and heads), $r = 8$
	parameters	$D = 438{,}784$ (UniDepthV2)
Adaptation	optimizer	AdamW (Loshchilov & Hutter, 2019), learning rate 10^{-2}
	weight decay	as released
	steps	10 or 40 (LoRA-TTO: 40)
	seeds	five for zero-initialized runs; SWS is deterministic
SWS	direction	rank-r truncated SVD of G_l, scaled by $\|W_l\|_F$ (8)
	amplitude	(9), 2 secant refinements; 1 backward and 4 forward passes
SRP	graph, affinity	four-neighborhood at full resolution; ranks (5), no bandwidth
	parametric fit	$[1, \hat{f}^2, \log \hat{f}]$ (7), double precision, $\hat{f}$ clamped to its range on $\mathcal{M}$
	solver	Jacobi-preconditioned CG, cascadic, 4×120 iterations; CUDA graph
Hardware	timing	one NVIDIA RTX 4090

Determinism. The forward pass is bitwise reproducible, and backward reductions are made deterministic, which the seed-variance measurements of Tab. 9 require: without it, repeated runs with the same seed would differ at the level being measured.

D ADDITIONAL RESULTS

All four metrics. Tab. 4 repeats Tab. 1 with AbsRel and δ_1. AbsRel tracks MAE closely everywhere, and δ_1 is near saturation on every benchmark except DDAD, where the frozen backbone's scale error is large enough to push pixels past the threshold. The ordering of the configurations is the same as in Tab. 1; in particular, the 10-step configuration matches or beats LoRA-TTO on all four metrics on every benchmark.

Table 4: **All four metrics for the rows of Tab. 1.** Same runs and 6,479 images; the step count is in parentheses.

Benchmark	Method (steps)	MAE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑
iBims-1	+ LoRA-TTO (Seo et al., 2026) (40)	0.041	0.146	0.011	0.994
	+ **SRP** *(ours)* (0)	0.040	0.162	0.011	0.992
	+ **SWS** + **SRP** *(ours)* (10)	0.041	0.144	0.011	0.994
	+ LoRA-TTO + SRP *(ours)* (40)	0.037	0.146	0.010	0.994
	+ SWS + SRP *(ours)* (40)	0.037	0.146	0.010	0.994
VOID	+ LoRA-TTO (Seo et al., 2026) (40)	0.136	0.505	0.026	0.982
	+ **SRP** *(ours)* (0)	0.152	0.539	0.030	0.975
	+ **SWS** + **SRP** *(ours)* (10)	0.134	0.504	0.025	0.982
	+ LoRA-TTO + SRP *(ours)* (40)	0.132	0.503	0.023	0.982
	+ SWS + SRP *(ours)* (40)	0.130	0.501	0.022	0.983
NYUv2	+ LoRA-TTO (Seo et al., 2026) (40)	0.054	0.125	0.019	0.992
	+ **SRP** *(ours)* (0)	0.055	0.142	0.020	0.989
	+ **SWS** + **SRP** *(ours)* (10)	0.054	0.125	0.019	0.992
	+ LoRA-TTO + SRP *(ours)* (40)	0.050	0.123	0.017	0.992
	+ SWS + SRP *(ours)* (40)	0.049	0.122	0.017	0.992
KITTI	+ LoRA-TTO (Seo et al., 2026) (40)	0.438	1.372	0.023	0.993
	+ **SRP** *(ours)* (0)	0.340	1.448	0.017	0.991
	+ **SWS** + **SRP** *(ours)* (10)	0.321	1.301	0.016	0.993
	+ LoRA-TTO + SRP *(ours)* (40)	0.318	1.325	0.015	0.993
	+ SWS + SRP *(ours)* (40)	0.310	1.329	0.015	0.994
DDAD	+ LoRA-TTO (Seo et al., 2026) (40)	2.002	5.937	0.061	0.962
	+ **SRP** *(ours)* (0)	1.480	6.325	0.041	0.968
	+ **SWS** + **SRP** *(ours)* (10)	1.386	5.494	0.037	0.971
	+ LoRA-TTO + SRP *(ours)* (40)	1.399	5.613	0.037	0.972
	+ SWS + SRP *(ours)* (40)	1.341	5.695	0.036	0.974

Paired statistics for the 10-step configuration. Sec. 5.2 gives paired statistics for the equal-budget row; the configuration we lead with receives the same treatment here. Averaging per-benchmark relatives, SWS at 10 steps read out through SRP changes MAE by -15.1% over the core and -12.3% over all five; pooled per image it gives -25.8% ($[-26.2, -25.5]$, $p < 10^{-4}$), improving 94.2% of the core's images. The pooled figure is larger for an arithmetic reason: DDAD is 69% of the core pool and has the largest gain. Weighting the per-benchmark relatives by image count already gives -26.2% against the pooled -25.8%, and -26.7% against -26.5% for the equal-budget row, so the imbalance accounts for essentially the whole difference. We therefore quote the unweighted per-benchmark figure as the headline throughout.

Win rates per metric. For SRP added to the 40-step run, MAE improves on 100.0% of images over the core and 99.9% over all five, and RMSE on 96.4% and 95.6%. The two smallest benchmark-level gains are VOID RMSE (0.33%) and iBims-1 RMSE (0.47%). At 10 steps the rates are 90.6% on MAE and 80.7% on RMSE over all five. RMSE is lower than MAE in both cases because a squared metric is priced by the pixels the field reaches least well, and on a minority of images that tail moves the wrong way even where the benchmark mean improves; 288 of the 6,479 images end with a higher RMSE at the full budget, most often on VOID (11% of its images).

Density and remaining error. One might suspect that the gain is purely a density effect. Over the core, the Spearman correlation between a sample's relative improvement and its measurement

density is $+0.34$ ($p < 10^{-4}$): since improvements are negative, the gain shrinks as measurements get denser. The correlation with the baseline's error on that sample is stronger, at -0.50: the gain is largest where the most error remains, whatever the density. This is the pattern Sec. 3.2 predicts, since the choice among minimizers is present on every sample, not only on sparse ones.

A second backbone. Tab. 5 repeats the leading rows on UniDepthV2 and MoGe-2, which differ both in head (a pixel decoder against a point-map head) and in how they obtain metric scale. Nothing is transferred and nothing is retuned: (8) reads each model's own gradient at its own frozen weights, and (9) is re-solved on its own ray. UniDepthV2 is re-run on the same 250 images as MoGe-2, so the two blocks form one comparison. At equal budget, SWS + SRP changes MAE against each backbone's own 40-step LoRA-TTO by -20.4% and -15.6% over the core (-16.7% and -12.8% over all five), wins all 4 driving cells on both, and is behind on 2 and 0 of the 6 indoor cells. With no backward pass, SRP changes MAE against a full 40-step adaptation of the same model by -21.3% and -24.0% on KITTI and by -27.3% and -25.4% on DDAD. At 10 steps the driving gains remain (-25.4% and -31.4% on UniDepthV2, -22.7% and -23.0% on MoGe-2), while 3 and 2 of the 6 indoor cells move the wrong way, by at most $+1.1\%$ (VOID RMSE) and $+0.8\%$ (VOID RMSE). On these subsets, the 10-step point trades a small indoor cost for large driving gains, and the equal-budget row removes the indoor cost.

Table 5: **The leading rows of Tab. 1 on two frozen backbones.** The adapter sits in whatever follows each frozen encoder, and nothing is retuned. **Bold and underlined** marks the best value in each column of each block. Absolute errors are comparable within a block but not across blocks, because each model predicts on its own grid. Both blocks use the same 250 images (50 iBims-1, 50 VOID, 50 NYUv2, 50 KITTI and 50 DDAD).

		iBims-1		VOID		NYUv2		KITTI		DDAD	
Method	Steps	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
UniDepthV2 (Piccinelli et al., 2026)											
frozen decoder + scale & shift	0	0.099	0.226	0.262	0.739	0.170	0.282	1.518	3.142	6.230	14.521
+ LoRA-TTO	40	0.043	0.154	0.208	**0.675**	0.054	0.128	0.453	1.406	2.057	6.140
+ **SRP** *(ours)*	**0**	0.042	0.170	0.216	0.696	0.055	0.147	0.356	1.504	1.496	6.486
+ **SWS** + **SRP** *(ours)*	**10**	0.042	**0.150**	0.209	0.682	0.053	0.129	0.338	**1.353**	1.411	**5.650**
+ SWS + SRP *(ours)*	40	**0.039**	0.156	**0.204**	0.676	**0.049**	**0.127**	**0.325**	1.377	**1.358**	5.894
MoGe-2 (Wang et al., 2025)											
frozen decoder + scale & shift	0	0.093	0.204	0.262	0.734	0.138	0.245	1.243	2.720	3.717	9.064
+ LoRA-TTO	40	0.042	0.148	0.206	0.670	0.055	0.122	0.439	1.421	1.812	6.109
+ **SRP** *(ours)*	**0**	**0.037**	0.155	0.212	0.688	**0.050**	0.132	0.333	1.395	**1.351**	5.869
+ **SWS** + **SRP** *(ours)*	**10**	0.040	**0.142**	0.206	0.676	0.052	0.120	0.339	**1.357**	1.395	**5.762**
+ SWS + SRP *(ours)*	40	**0.037**	**0.142**	**0.202**	**0.668**	0.051	**0.119**	**0.332**	1.377	1.448	6.081

The budget axis. Tab. 6 sweeps the adaptation budget for both initializations, with and without the read-out. **The read-out substitutes for most of the adaptation**: 2 adaptation steps read out through SRP reach 92% of the 40-step baseline's MAE over the core, against 227% without it. The gain is concentrated on the driving benchmarks, and 2 steps leave the RMSE tail behind almost everywhere; closing it is what further steps buy. **The warm start makes the short budget usable**: at 10 steps with the read-out, the zero initialization beats the baseline in 5 of the 10 benchmark–metric cells, while SWS matches or beats it in all 10; without the read-out, SWS reaches the baseline's full-budget accuracy in 30 steps. The steps a zero-initialized run spends leaving the degenerate point of (C3) are a fixed cost, so removing them matters most when the budget is tightest.

Latency. Tab. 7 breaks the latency down by component and resolution. Solving (4) by plain conjugate gradients to convergence costs 1,425 ms at VGA; the cascadic cycle comes within 1.1% of that field (Tab. 11), and capturing it saves a further factor of 5–10. The uncaptured solve costs 0.45 ms per iteration at every resolution we tried, from 0.3 to 2 megapixels; a cost that does not grow with the problem is launch overhead, which is exactly what capture removes. At the adaptation-free point the whole pipeline runs at 9.2 frames per second at VGA and 1.0 at FHD.

Table 6: **The budget axis.** Both initializations at eight adaptation budgets, with and without the read-out; rows differ only in the initialization, the number of steps and the read-out. Bold rows mark the baseline's shipped budget and the configuration we lead with. 895 images, up to 200 per benchmark.

	iBims-1		VOID		NYUv2		KITTI		DDAD	
Init. / steps	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
the adaptation alone										
LoRA-TTO, 2	0.092	0.192	0.220	0.720	0.155	0.250	0.961	2.217	3.591	9.926
LoRA-TTO, 5	0.076	0.170	0.207	0.694	0.117	0.192	0.890	1.951	3.518	7.955
LoRA-TTO, 10	0.061	0.155	0.194	0.675	0.089	0.157	0.692	1.661	2.815	6.986
LoRA-TTO, 15	0.055	0.152	0.186	0.671	0.075	0.140	0.588	1.506	2.554	6.501
LoRA-TTO, 20	0.049	0.149	0.182	0.668	0.067	0.133	0.534	1.445	2.366	6.250
LoRA-TTO, 25	0.046	0.147	0.178	0.663	0.062	0.130	0.497	1.411	2.276	6.167
LoRA-TTO, 30	0.044	0.147	0.177	0.662	0.059	0.127	0.466	1.373	2.148	6.043
LoRA-TTO, 40	**0.041**	**0.147**	**0.174**	**0.660**	**0.055**	**0.125**	**0.429**	**1.353**	**1.978**	**5.909**
SWS, 2	0.094	0.183	0.205	0.693	0.124	0.201	0.948	2.069	3.248	7.649
SWS, 5	0.084	0.186	0.194	0.678	0.099	0.169	0.716	1.688	2.843	6.987
SWS, 10	0.058	0.154	0.185	0.666	0.078	0.144	0.583	1.496	2.403	6.367
SWS, 15	0.053	0.151	0.181	0.663	0.068	0.135	0.519	1.428	2.214	6.092
SWS, 20	0.048	0.149	0.178	0.662	0.062	0.129	0.477	1.382	2.062	5.947
SWS, 25	0.045	0.149	0.176	0.661	0.058	0.127	0.446	1.354	1.976	5.917
SWS, 30	0.043	0.148	0.174	0.659	0.056	0.125	0.423	1.346	1.879	5.865
SWS, 40	0.040	0.147	0.172	0.658	0.053	0.124	0.395	1.339	1.765	5.850
+ *SRP*										
no adaptation	0.040	0.162	0.193	0.693	0.056	0.143	0.339	1.435	1.484	6.404
LoRA-TTO, 2	0.046	0.151	0.187	0.682	0.058	0.131	0.326	1.315	1.454	6.190
LoRA-TTO, 5	0.045	0.149	0.178	0.668	0.061	0.131	0.333	1.313	1.471	5.694
LoRA-TTO, 10	0.042	0.146	0.175	0.663	0.057	0.127	0.329	1.306	1.451	5.581
LoRA-TTO, 15	0.041	0.147	0.173	0.663	0.055	0.124	0.325	1.299	1.436	5.528
LoRA-TTO, 20	0.040	0.146	0.173	0.662	0.054	0.124	0.323	1.302	1.427	5.536
LoRA-TTO, 25	0.039	0.145	0.171	0.660	0.053	0.124	0.321	1.308	1.416	5.538
LoRA-TTO, 30	0.038	0.145	0.171	0.659	0.052	0.123	0.319	1.307	1.410	5.566
LoRA-TTO, 40	0.037	0.146	0.170	0.658	0.051	0.123	0.316	1.312	1.396	5.608
SWS, 2	0.046	0.149	0.176	0.665	0.060	0.129	0.329	1.311	1.420	5.675
SWS, 5	0.041	0.151	0.174	0.662	0.057	0.127	0.324	1.294	1.408	5.589
SWS, 10	**0.041**	**0.144**	**0.171**	**0.657**	**0.055**	**0.125**	**0.320**	**1.290**	**1.387**	**5.530**
SWS, 15	0.040	0.147	0.171	0.658	0.053	0.124	0.318	1.299	1.376	5.526
SWS, 20	0.039	0.146	0.170	0.658	0.052	0.123	0.314	1.294	1.366	5.553
SWS, 25	0.038	0.147	0.171	0.658	0.051	0.123	0.312	1.297	1.356	5.593
SWS, 30	0.038	0.147	0.170	0.657	0.050	0.123	0.310	1.304	1.347	5.621
SWS, 40	0.037	0.146	0.169	0.657	0.050	0.122	0.308	1.310	1.338	5.696

Table 7: **Latency by input resolution.** Milliseconds per image, median (protocol in App. C). "Steps" is the number of adaptation steps per image. Indented rows decompose the row above; italic rows are variants of the deployed solve that we do not ship. The warm start's cost, one gradient evaluation and 4 forward passes, is charged once per image.

		latency (ms) by input resolution			
Method	Steps	VGA	SD	HD	FHD
Frozen forward pass	0	88	101	303	976
TestPromptDC (Jeong et al., 2025)	—	6,793	8,404	34,035	134,346
LoRA-TTO (Seo et al., 2026)	40	4,343	4,869	12,361	30,354
SWS *(ours)*	40	4,497	5,062	13,197	32,632
SRP only *(ours)*	**0**	109	123	331	1,024
affinity (5)		1.4	1.4	1.9	4.4
parametric fit (7)		1.3	1.3	1.8	2.9
cascadic solve, captured		18.3	18.8	24.1	40.3
same, not captured		179	180	184	182
same, plain CG to convergence		1,425	1,431	1,429	1,431
SWS + SRP *(ours)*	**10**	1,728	1,916	4,551	11,149
SWS + SRP *(ours)*	40	4,517	5,083	13,225	32,680

E Empirical Verification of the Analysis

The statements of Sec. 3 are proved in App. A. What remains to check is that the objective the code minimizes is the one analyzed, and that the effects the three consequences predict are large enough to matter. Tab. 8 does the first on every sample of the diagnostic subset; the figures below do the second.

Table 8: **The closed form and its consequences.** Measured on the 50-image diagnostic subset (range over samples). First block: the implemented objective against (2), and the flat manifolds of (C2), the factorization one over scalar rescalings and over the full group $\mathrm{GL}(r)$. Second block: the degenerate initialization (C3), with A and B the up- and down-projections. Third block: the data scale (C1), the under-determination of Sec. 3.2, and the submersion check (App. E).

Property	Quantity	Measured over the sample set
Closed form (2)	max. rel. deviation from the implemented loss	8.1×10^{-8} – 1.7×10^{-6}
Envelope identity	$\cos(\nabla\mathcal{L}_{\text{detached}}, \nabla\mathcal{L}_{\text{exact}})$	0.999984 – 1.000000
Output-affine invariance	max. rel. deviation of $\mathcal{L}$	0 – 1.1×10^{-6}
Factor rescale invariance	max. rel. deviation of $\mathcal{L}$	1.4×10^{-6} – 8.0×10^{-4}
the full gauge group	the same, under $(A, B) \mapsto (AM, M^{-1}B)$	1.7×10^{-5} – 5.0×10^{-3}
Degenerate init. (C3)	fraction of samples with $\nabla_B\mathcal{L} \equiv 0$	50/50
	fraction of negative Ritz values	0.47 – 0.50
	$\|\nabla_A\mathcal{L}\|/\|\nabla_B\mathcal{L}\|$ near init.	3.2×10^{4} – 2.4×10^{5}
Data scale (C1)	$\mathcal{L}$ at initialization	1.0×10^{-3} – 6.4×10^{2}
	$1 - \rho^2_{\mathcal{M}}$ at initialization	3.6×10^{-3} – 4.9×10^{-1}
Under-determination	measured pixels m	419 – 17,923
	$1 - \rho^2_{\mathcal{M}}$ after adaptation	3.1×10^{-5} – 1.8×10^{-1}
	affine share of the induced output change	1.4% – 99.5%
Submersion (Sec. 3.2)	random 64-row block of $\partial f_\theta\|_{\mathcal{M}}/\partial\theta$ at full rank	150/150
	$\sigma_{\min}/\sigma_{\max}$ of that block	1.1×10^{-4} – 8.6×10^{-3}

The implemented objective. The implemented loss matches (2) to within 1.7×10^{-6} relative deviation, the detached and exact gradients have cosine at least 0.99998, and both flat manifolds of (C2) hold to floating-point accuracy: the output-affine invariance to 1.1×10^{-6} and the factorization invariance to 5×10^{-3} relative deviation over the full group $\mathrm{GL}(r)$—diagonal, permutation, orthogonal and general M with condition numbers up to 10^2, drawn independently per adapted layer—not only over scalar rescalings (Tab. 8).

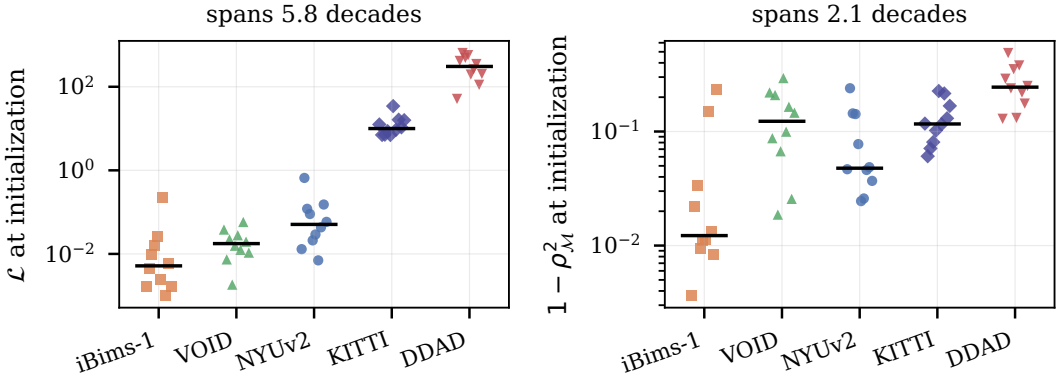

Figure 4: **The objective at initialization, per sample and benchmark.** Bars mark medians. *Left*: the objective. *Right*: the same 50 samples divided by $\mathrm{Var}_{\mathcal{M}}(y)$, i.e., the dimensionless factor $1-\rho^2_{\mathcal{M}}$ of (2).

(C1) The data scale. Fig. 4 shows the objective at initialization before and after dividing out $\mathrm{Var}_{\mathcal{M}}(y)$: the span shrinks from about six orders of magnitude to about two. The remaining spread is real—indoor scenes are closer to the frozen prediction than driving scenes—but small enough that one dimensionless amplitude serves every benchmark (App. G).

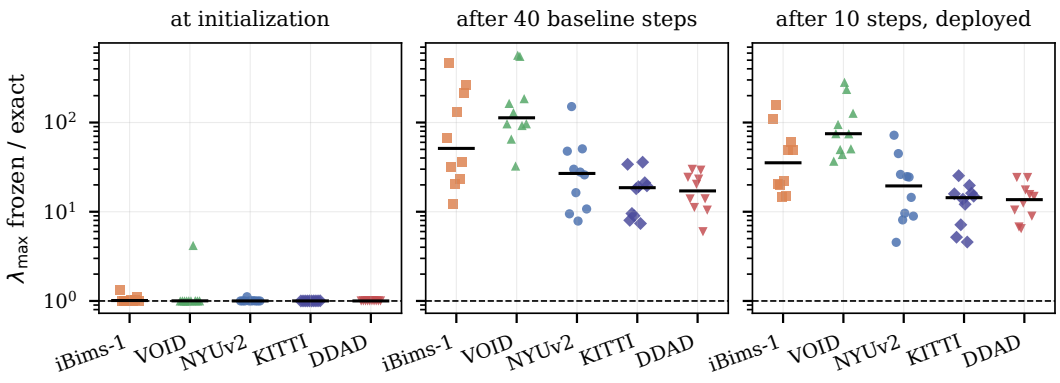

Figure 5: **Largest Hessian eigenvalue with the scale and shift held constant, divided by the same quantity with the alignment differentiated.** Bars mark medians over 50 images, at the initialization, after the baseline's 40 steps and after the 10 warm-started steps we deploy.

(C2) Curvature must be computed with the alignment differentiated. Fig. 5 measures this at three points on the trajectory. The two estimates agree at the initialization, where the adapter's effect is exactly zero; after the baseline's 40 steps and after our 10 warm-started steps they diverge alike, by a median factor of 20 (range 5–282) on the deployed 10-step run. A frozen alignment reports curvature along directions in which (2) is exactly flat, so any curvature-based diagnosis of this objective, including of our own runs, must differentiate the alignment. All curvature quantities here use Hessian-vector products with the alignment differentiated, extremal eigenvalues by Lanczos iteration and traces by Hutchinson probes, restricted to the parameters that can influence the depth output; in this backbone some adapter tensors feed only the confidence head and would otherwise dilute every spectral quantity.

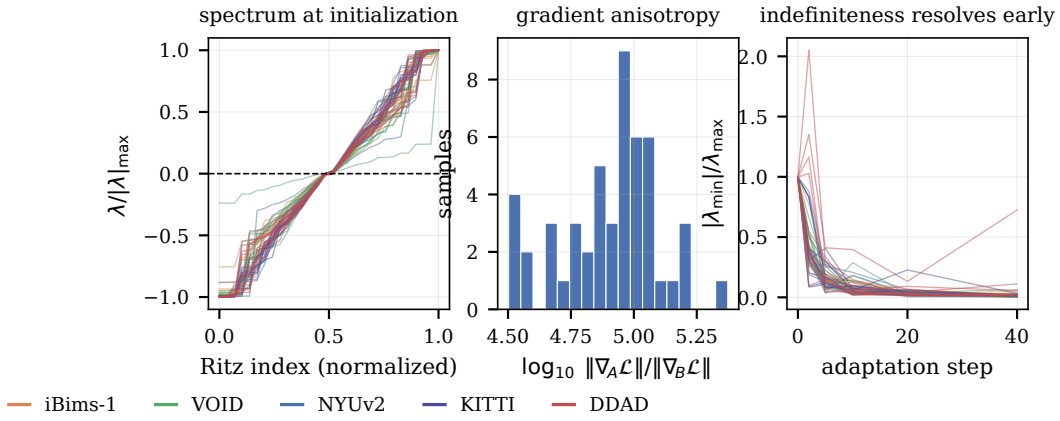

Figure 6: **The Hessian at the zero initialization.** Computed over the 50-image diagnostic subset. *Left*: Ritz values, normalized by the largest magnitude, one curve per sample. *Middle*: the ratio of the two factors' gradient norms a millionth of the way along the trajectory (at the initialization itself the denominator is exactly zero). *Right*: $|\lambda_{\min}|/\lambda_{\max}$ along the run.

(C3) The degenerate initialization. Fig. 6 shows (C3) in three views. The Ritz spectrum at the initialization is close to antisymmetric on every sample (between 47% and 50% of the Ritz values are negative), as expected when the cross block dominates the Hessian; the gradient-norm ratio between the two factors is four to five orders of magnitude a millionth of the way along the trajectory; and the indefiniteness resolves within a few steps, steps that a short budget can ill afford. Tab. 9 shows the same effect through the seed: the spread over five adapter initializations grows with the depth range, because the seed selects the escape direction from the degenerate point. SWS overwrites both factors, so its spread is exactly zero.

Table 9: **Standard deviation of MAE (m) over five adapter seeds.** Rows with SWS are exactly zero because the warm start overwrites both factors.

Method	iBims-1	VOID	NYUv2	KITTI	DDAD
LoRA-TTO	0.0010	0.0032	0.0013	0.0198	0.1073
SWS (either read-out)	0.0000	0.0000	0.0000	0.0000	0.0000

The submersion hypothesis. The codimension of Prop. A.5 is $m - 2$ only where $\theta \mapsto f_\theta|_{\mathcal{M}}$ has full row rank. The Jacobian is $m \times D$ with $D = 438{,}784$ and m up to $18{,}966$, so neither forming it nor computing its smallest singular value is affordable per sample, but a random 64-row block is. Full rank of every such block is necessary for the hypothesis but not sufficient: a rank drop would falsify it, and full rank at healthy conditioning is evidence for it. We measure at three points on the trajectory—the frozen model, the 10-step state we deploy and the 40-step baseline endpoint—and every block is full rank ($150/150$ sample–point pairs), with $\sigma_{\min}/\sigma_{\max} \geq 1.1 \times 10^{-4}$ throughout.

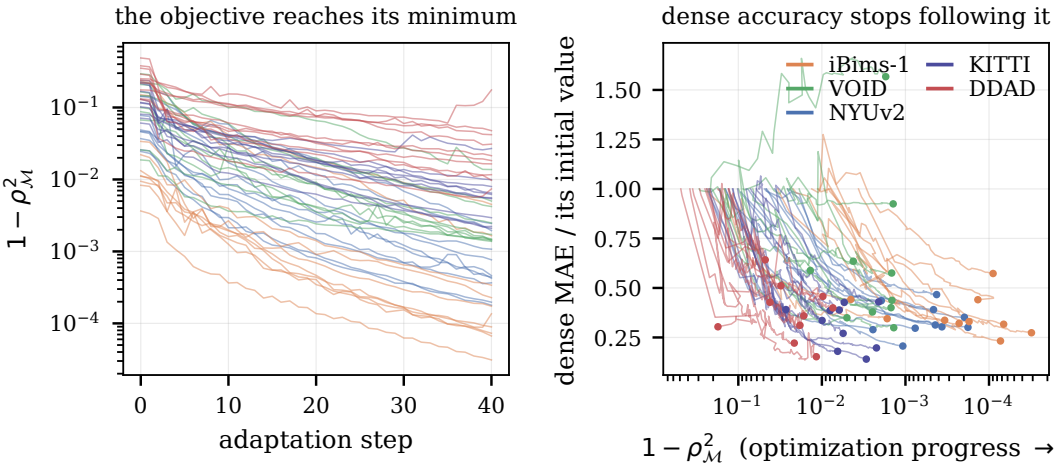

Figure 7: **The objective along the run.** One curve per sample over the 50-image diagnostic subset. *Left*: the objective against the step count. *Right*: dense accuracy against the objective, so progress runs from left to right.

Drift along the minimizer set. Fig. 7 is the measurement behind Sec. 3.2. The objective saturates early, yet dense accuracy continues to improve: the run keeps moving along $\Theta^\star$, and the drift changes the unmeasured pixels. This is why the adaptation budget matters even though the objective is essentially solved, and why choosing the point directly, as SRP does, is much cheaper than buying the drift with more steps.

F Ablations of the Propagation

One design decision at a time. Tab. 10 removes one decision from SRP at a time. Two rows put a constant back and two do not. Writing the measurements as a penalty costs 18% on the frozen prediction even with μ tuned on the calibration set; replacing (5) by $\exp(-\|\Delta I\|/\sigma)$ with a tuned σ costs 12%; and both reintroduce a per-benchmark constant in data units, which (C1) says does not transfer. A uniform affinity costs 27%, more than any other single decision, and dropping the parametric term costs 13% of RMSE on DDAD. After 40 steps all variants are close to full SRP: these decisions carry information a long way, and after adaptation the residual is small and local.

Image structure and density. The uniform-affinity row asks what guided filtering (He et al., 2010) asks: what does image structure buy? Sorting samples by measurement density sharpens the answer. On the same runs as Tab. 10, removing the image still costs $+3.1\%$ of MAE on the three sparsest quartiles after adaptation, and nothing (-0.1%) on the densest. Image structure is what lets a residual travel, so it matters where measurements are sparse and the residual is large. We do not gate the affinity on density, since the gate would be one more constant to choose.

Table 10: **One design decision of SRP removed at a time.** Each variant is evaluated with no adaptation and after 40 steps. "Const." counts the values each variant must choose per benchmark; each variant keeps the rest of SRP fixed. 895 images, stratified across the five benchmarks.

		iBims-1		VOID		NYUv2		KITTI		DDAD	
Variant	Const.	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
no adaptation											
SRP *(ours)*	0	0.040	0.162	0.193	0.693	0.056	0.143	0.339	1.435	1.484	6.404
measurements as a penalty, tuned μ	1	0.049	0.171	0.208	0.717	0.064	0.156	0.355	1.504	1.961	7.221
affinity by bandwidth, tuned σ	1	0.046	0.167	0.198	0.698	0.061	0.148	0.367	1.520	1.666	6.567
uniform affinity (no image, no depth)	0	0.055	0.174	0.201	0.702	0.076	0.166	0.377	1.488	1.795	6.591
without the parametric term (7)	0	0.040	0.163	0.190	0.691	0.055	0.143	0.342	1.485	1.572	7.392
after 40 adaptation steps											
SRP *(ours)*	0	0.037	0.146	0.170	0.658	0.051	0.123	0.316	1.312	1.396	5.608
measurements as a penalty, tuned μ	1	0.038	0.146	0.171	0.659	0.051	0.123	0.316	1.311	1.420	5.620
affinity by bandwidth, tuned σ	1	0.038	0.146	0.170	0.658	0.051	0.123	0.318	1.317	1.426	5.626
uniform affinity (no image, no depth)	0	0.038	0.146	0.170	0.658	0.052	0.124	0.316	1.305	1.440	5.599
without the parametric term (7)	0	0.037	0.146	0.170	0.659	0.051	0.123	0.316	1.313	1.396	5.608

The solver budget is monotone. Tab. 11 solves (4) four ways: plain conjugate gradients at 400 and 3,200 iterations, the cascadic cycle at the deployed fixed budget, and the same cycle stopped on a residual tolerance. They bracket the same field, and the deployed budget is within 1.1% of convergence at a fraction of the cost. There is no interior optimum to find, so the budget is a cost choice; and only the fixed-budget cycle has a loop length set by the resolution alone, so only it can be captured as a graph.

Table 11: **One problem, four solvers.** Every row solves (4) on the frozen prediction with the same affinity; the last row adds (7) to the deployed solve, which is SRP. 6,479 images.

		iBims-1		VOID		NYUv2		KITTI		DDAD	
Solver	Iters	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
plain CG, 400 iterations	400	0.040	0.162	0.148	0.533	0.054	0.142	0.343	1.498	1.661	8.040
plain CG, 3,200 iterations	3,200	0.039	0.162	0.147	0.530	0.054	0.142	0.342	1.482	1.514	6.724
cascadic, residual tolerance 10^{-4}	adaptive	0.039	0.162	0.147	0.531	0.054	0.142	0.342	1.483	1.514	6.726
cascadic, 4×120 **(deployed)**	480	0.040	0.163	0.149	0.534	0.054	0.142	0.343	1.491	1.568	7.271
the same, + (7) *(SRP)*	480	0.040	0.162	0.152	0.539	0.055	0.142	0.340	1.448	1.480	6.325

Reach. Tab. 12 is the measurement behind Sec. 4.1 and is reported per benchmark because pooling destroys it. The maximum principle predicts not that the solve is worse than adaptation but that the margin grows with distance, and it does so on exactly two benchmarks. On NYUv2 the ratio falls slightly with distance, from 1.31 to 1.14; on iBims-1 it rises mildly, from 0.97 to 1.30; and on VOID it stays near parity throughout, even though VOID has the largest share of distant pixels of any benchmark (over a quarter beyond 16 pixels): distance alone does not make the ceiling bind. On DDAD the solve is better than parity within 4 pixels and 5.7× worse between 16 and 64 (5.2–6.2); (7) cuts that group to 2.3× (2.1–2.5), a factor of 2.5 with disjoint intervals, and improves the nearer groups as well. KITTI shows the same shape at less than half the size (2.4× in the same group before (7)), as its measurement pattern predicts: 96% of its evaluated pixels lie within 4 pixels of a measurement. A better graph or a longer solve could not supply this, since the converged field is still bounded by (6); the parametric term is not.

Where the prior is placed. The solve (4) applies a smoothness prior to the correction rather than to the depth. Tab. 13 tests the assumption behind this directly, with no accuracy metric in the loop, by measuring how strongly each field concentrates its total variation at the ground-truth depth discontinuities; the correction field scored is the one SRP produces on the deployed configuration. Indoors, ground-truth depth concentrates its variation at its discontinuities far more than any computed field (18.5–65.4 against at most 6.6), so a prior that is smooth away from edges is a poor description of depth. Among the three computed fields, the correction is the least concentrated on four of the five benchmarks and over the pooled set (3.31 against 4.56 and 5.43), and on DDAD it is level with the frozen prediction (3.99 against 3.87): it is the field for which the prior fits best, which is why (4) is applied to it. The separation is narrowest on the driving benchmarks, whose ground truth is itself least concentrated at its discontinuities (5.0–5.6).

How many parametric terms. Tab. 14 sweeps nested bases on the frozen prediction, scoring each by dense error and by the distance decomposition of Tab. 12, so that reach and generalization are

Table 12: **Error against distance to the nearest measurement.** Per benchmark, the first row is the share of evaluated pixels in each group; the others give the summed squared error of SRP on the frozen prediction, with and without (7), divided by that of a 40-step LoRA-TTO run on the same group (bold: above 2). Groups with fewer than 10,000 evaluated pixels print the count instead. 2,767 images, 95–768 per benchmark.

		distance to the nearest measurement (px)			
Benchmark		0–4	4–16	16–64	> 64
iBims-1	evaluated pixels	14.2%	78.1%	7.6%	*892 px*
	SRP without (7)	0.97	1.14	1.30	—
	SRP	0.96	1.13	1.27	—
VOID	evaluated pixels	11.6%	60.5%	25.4%	2.6%
	SRP without (7)	1.06	1.02	1.04	1.07
	SRP	1.06	1.01	1.04	1.09
NYUv2	evaluated pixels	27.5%	71.9%	0.6%	0
	SRP without (7)	1.31	1.26	1.14	—
	SRP	1.31	1.25	1.14	—
KITTI	evaluated pixels	96.0%	4.0%	0.037%	*1 px*
	SRP without (7)	1.13	1.24	**2.36**	—
	SRP	1.07	1.15	1.97	—
DDAD	evaluated pixels	39.8%	58.5%	1.7%	*672 px*
	SRP without (7)	0.95	1.46	**5.67**	—
	SRP	0.86	1.19	**2.30**	—

Table 13: **Concentration $C(u)$ of a field's total variation at ground-truth discontinuities.** The mean of $\|\nabla u\|_1$ over the 5% of valid pixels with the largest ground-truth gradient, divided by the mean over the rest. *All* pools every sample. 50 images.

Field	iBims-1	VOID	NYUv2	KITTI	DDAD	All
Ground-truth depth	55.74	65.36	18.54	5.57	5.04	14.60
Frozen prediction	6.62	3.17	5.47	4.95	3.87	4.56
Adapted prediction $\hat{f}$	6.11	3.35	6.28	5.61	5.50	5.43
Correction field $\delta^\star$	3.79	2.09	2.78	3.89	3.99	3.31

separate columns of one run. More terms reach further: the 16–64 group on DDAD falls from $2.09\times$ LoRA-TTO's squared error at three terms to $1.86\times$ at four, and no further ($1.87\times$ at six to fourteen terms). More terms also generalize worse: dense error rises monotonically from 93.4% of LoRA-TTO's at three terms to 95.8% at fourteen, because every added coefficient is fitted on a measured set that is a biased sample of the scene. The balance tips between three and four terms; we keep three, since the four-term gain is 0.2 in a group holding under 2% of DDAD's pixels, against a dense cost paid on all of them. A trend in *position*, as in regression kriging, gives 94.0% dense error and $6.16\times$ in the 16–64 group: a trend on the grid is another field on the grid and says nothing about a far pixel that the solve does not. An IRLS-reweighted fit is worse on both columns, as expected when residuals are structured rather than contaminated. Re-estimating the affinity from the corrected depth and solving again changes nothing—the second solve reproduces the first to seven digits—because the first solve already matches every measurement exactly, so the second receives a residual that is zero on $\mathcal{M}$: a property of the constrained form that a penalized solve would not have.

Table 14: **Size of the parametric basis.** Measured on the frozen prediction with everything else in SRP fixed. "MAE / LoRA-TTO" is the mean over the five benchmarks of dense error relative to a 40-step LoRA-TTO run; the last column is the squared-error ratio in DDAD's 16–64 px group (Tab. 12). x, y are pixel coordinates on $[-1, 1]$. 495 images, up to 100 per benchmark.

Parametric term	Terms	MAE / LoRA-TTO	16–64 px / LoRA-TTO
$[1, \hat{f}^2, \log \hat{f}]$ (7)**, deployed**	3	93.4%	2.09
$+ \hat{f}$	4	93.9%	1.86
$+ x, y$	6	94.2%	1.87
$+ x^2, y^2, xy$	9	94.9%	1.87
$+ \hat{f}\,\{x, y, x^2, y^2, xy\}$	14	95.8%	1.87
$[1, x, y]$, a trend in *position*	3	94.0%	6.16
$+ x^2, y^2, xy$	6	94.4%	6.18
three terms, IRLS-reweighted fit	3	93.6%	4.58
three terms, affinity re-estimated and re-solved	3	93.4%	2.09

G Ablations of the Warm Start

Components. Tab. 15 separates SWS into its parts at both budgets. The LoRA-GA-style row is the controlled contrast—same gradient, factorization, optimizer and budget, with the step canceled—and aligning the initialization this way is worth 3.9% at 40 steps and 8.6% at 10. Taking the step, scaled per layer and with its computed length, adds 3.4% and 5.3%, for totals of 7.3% and 13.9%. Cutting the budget to a quarter roughly doubles the value of the alignment: a long run finds a good subspace by itself and a short one does not. A random nonzero up-projection of the same magnitude is neutral (Tab. 17), so what matters is where the starting point is, not that it is nonzero.

Table 15: **Components of the warm start.** The indented row adds to the one above. Columns: median final MAE relative to LoRA-TTO, mean seed standard deviation and the fraction of samples ending worse than the frozen backbone (over the core). The LoRA-GA-style row cancels the step. 247 images, stratified across the five benchmarks.

	40 steps			10 steps		
Variant	MAE / LoRA-TTO	seed sd	fail	MAE / LoRA-TTO	seed sd	fail
LoRA-TTO (no warm start)	1.000	0.0225	0.0%	1.000	0.0488	0.5%
Spectral step, no per-layer scale	0.974	0.0000	0.0%	0.913	0.0000	1.5%
+ per-layer scale $\|W_l\|_F$ (SWS)	0.927	0.0000	0.0%	0.861	0.0000	1.5%
LoRA-GA-style init (Wang et al., 2024) (step canceled)	0.961	0.0000	0.0%	0.914	0.0000	1.5%

The per-layer scale. Dropping $\|W_l\|_F$ from (8) costs 5.2% at 10 steps and 4.7% at 40. With the amplitude computed by (9), the absolute variant no longer diverges, since (9) scales its coefficient to whatever ray it is given, so the per-layer factor now buys accuracy rather than stability: it sets where the step lands. At 10 steps, every warm-started variant ends worse than the frozen backbone on 1.5% of samples against the baseline's 0.5%; at 40 steps no variant does. Every SWS variant has a seed standard deviation of exactly zero, against 0.049 for the baseline at 10 steps.

Why one τ can be shared. Fig. 8 sweeps a fixed amplitude. The curve is flat below a threshold and a cliff above it, and the collapse begins at the same τ on benchmarks whose depth scales differ by two orders of magnitude, which is what the per-layer $\|W_l\|_F$ provides. The curve is also asymmetric: the first-order model of (9) errs low, since the tangent at the origin underestimates the ray's maximizer, and low is the inexpensive side. At 10 steps the choice within the flat region is worth a few points; at 40 it is worth a fraction of one, because a longer run absorbs the initialization's error.

The amplitude rule and the refinement count. Tab. 16 compares (9) with greedy search and with one fixed τ, on 495 images drawn evenly from the five evaluation splits; the final configuration has no held-out set. The comparison gives the search every advantage. The grid we started from was too low—its upper half is selected on 1 of 495 samples, and 48% of samples select its lowest point—so we also report a grid placed on the useful range and scored on held-out pixels or on the training loss. The best grid and a single fixed value land within 0.7 points of each other, with no setting reliably best, and (9) with two refinements is better than the best search by 0.3 points at 10 steps and 0.6 at 40, for 4 forward passes against 6 and with nothing to choose. The only column in which a grid is

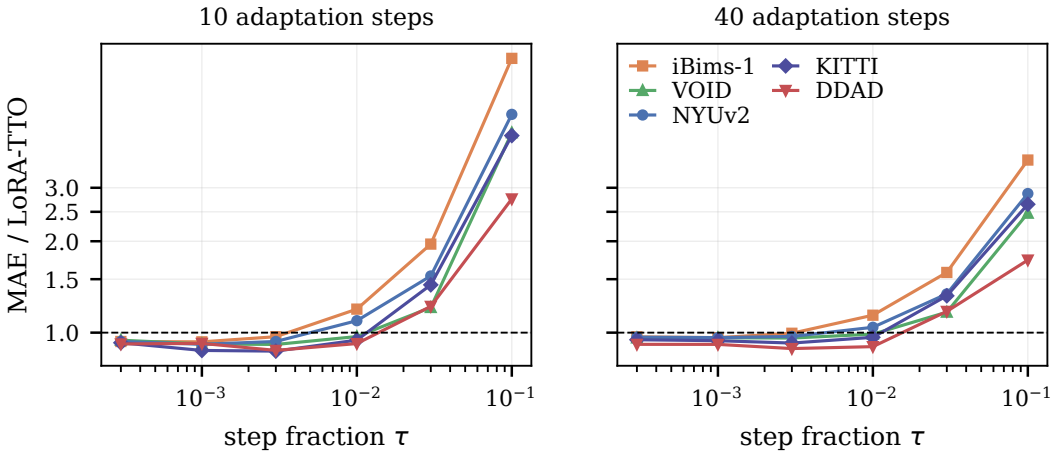

Figure 8: **Median MAE relative to LoRA-TTO at the same budget, against a fixed amplitude, the step fraction τ in (8).** One τ is shared by every sample, with no excursion cap. *Left*: 10 steps. *Right*: 40 steps. 250 images.

ahead, by 0.3 points, is the 10-step run without the read-out, which no reported configuration uses. Two refinements beat none and three, so the count has an interior optimum; unlike a constant in data units it is one dimensionless integer, fixed for all five benchmarks at once. A third re-solve lands closer to the ray's $\rho^2_{\mathcal{M}}$ maximizer yet is worse on dense error, because that maximizer is defined on the measured pixels; this is also why the count is fixed globally and never selected per sample.

Table 16: **How the amplitude is set.** Same direction (8) and recipe throughout; "Const." counts the values to choose and "Fwd." the forward passes spent. Entries are the mean over the five benchmarks of MAE relative to LoRA-TTO at the same budget (100 is parity). 495 images.

			10 steps		40 steps
How the amplitude is set	Const.	Fwd.	alone	+ SRP	+ SRP
greedy search, grid on the useful range	1	6	89.2	96.8	98.2
the wider grid we started from	1	7	88.7	96.8	98.5
one fixed τ, shared by every sample	1	1	89.3	97.4	98.8
(9), no refinement	0	1	92.5	96.9	98.0
+ **2 refinements** *(deployed)*	0	4	89.0	96.5	97.7
+ 3 refinements	0	5	89.3	96.7	97.9

H Design Alternatives That Do Not Help

Sec. 3.2 predicts that interventions improving the *optimization* should not help, because the optimization already succeeds, and neither should interventions regularizing in *parameter* space, because the dimension of $\Theta^\star$ is set by the number of unmeasured pixels. App. H.1 tests both classes; App. H.2 reports alternatives the analysis does not rule out but that a reader may consider. All use the backbone and budget of the main results, on the subsets stated in the captions.

H.1 Alternatives the analysis predicts should not help

Optimization and parameter-space regularization. Tab. 17 lists twelve interventions applied to LoRA-TTO. Ten are at or above parity: a warm-up with a cosine schedule, factor-asymmetric learning rates (Hayou et al., 2024) (the more aggressive ratio fails on every sample), sharpness-aware minimization (Foret et al., 2021), parameter averaging, early stopping on held-out measured pixels, a proximal term, two output trust regions and a nonzero random up-projection. The two below parity move along the exactly flat directions of (C2) and are within half a percent, the expected outcome for changes the objective cannot see. LoRA-TTO is also at its step-size ceiling (Tab. 19): one grid step higher diverges, and raising the learning rate *after* the warm start diverges at the same threshold, so the ceiling is a property of the objective and not of the starting point. The same table bounds the

effect of weight decay at about 2%; it is the one setting in which our runs could have differed from LoRA-TTO, and they do not.

Table 17: **Optimization and parameter-space interventions.** Each is applied to LoRA-TTO with everything else fixed. Columns: median final MAE relative to LoRA-TTO and the fraction of runs ending worse than the frozen backbone. 50 images, over the core.

Intervention	MAE / LoRA-TTO	fail
Warm-up + cosine schedule	1.079	0.0%
LoRA+, ratio 4 (Hayou et al., 2024)	1.316	13.3%
LoRA+, ratio 16 (Hayou et al., 2024)	15.127	100.0%
Sharpness-aware minimization (Foret et al., 2021)	1.002	0.0%
Parameter EMA over the last half of the run	1.017	0.0%
Early stopping on held-out measured pixels	1.086	0.8%
Proximal term toward the initialization	1.206	0.0%
Output trust region, 0.10	1.009	0.0%
Output trust region, 0.20	1.000	0.0%
Nonzero random up-projection	1.008	0.0%
Balanced factor norms	0.996	0.0%
Affine-free adapter effect	0.997	0.0%

The outer loss. The closed form holds because the inner alignment is the *exact* minimizer of a squared error. Tab. 18 replaces the loss with four alternatives, and each is worse: ℓ_1, Huber, ℓ_1+RMSE and SILog each leave at least one benchmark worse than no adaptation at all, and both robust losses fail on both driving benchmarks, where the measured set contains the returns a robust loss is designed to discount. None of them admits the closed form (2), so none of the analysis applies to them.

Table 18: **Outer loss in (1).** Median final MAE relative to the frozen backbone (> 1 is worse than no adaptation). 50 images.

Loss	iBims-1	VOID	NYUv2	KITTI	DDAD
MSE (default)	0.335	0.505	0.309	0.330	0.421
ℓ_1+RMSE	0.321	2.304	0.339	0.378	0.819
Huber	0.327	0.513	0.315	1.991	3.162
ℓ_1	0.484	3.852	0.535	2.509	3.387
SILog	0.570	1.691	1.355	6.267	3.552

Table 19: **Optimizer and learning rate of the inherited recipe.** Median final MAE relative to the frozen backbone, over the core. 50 images.

Optimizer, learning rate	MAE / frozen backbone
SGD, 10^{-1}	0.999
AdamW, 10^{-3}	0.402
AdamW, 3×10^{-3}	0.329
AdamW, 10^{-2} *(reference)*	0.306
Adam (no weight decay), 10^{-2}	0.312
AdamW, 3×10^{-2}	3.838

Restricting the adapter to a subspace. The trajectory looks low-dimensional—its first principal component carries most of the variance, and the merged weight update has stable rank near one—which suggests freezing the down-projection at the leading singular vectors of G_l and training only the up-projection. This reaches $1.15\times$ LoRA-TTO's MAE (median over the benchmarks). The subspace a run needs is not fixed: the initial subspace captures a median 0.41 of the final update's energy. (We do not quote principal angles: in an ambient space this large, the largest angle between two rank-8 subspaces is near 90° for almost any pair.) The optimizer rotates its own subspace during the run, which is why SWS uses the spectral step only as an initialization and then frees both factors.

Extrapolating the trajectory. If the path from the initialization to the endpoint were straight, moving further along it would keep improving. At $\theta_0 + 1.5(\theta_T - \theta_0)$, however, the objective is $79\times$ its value at θ_T (median over the benchmarks): the path along $\Theta^\star$ is curved, and its endpoint cannot be reached early by extrapolation.

H.2 Other alternatives

Folding the solve into the adapter. If the field the solve produces is the one we want, one could train the adapter to produce it and drop the second stage. Tab. 20 reports five training losses in three families, under one protocol. *Distillation* replaces the objective by agreement with the converged field, in the closed form of (2) so that a constant prediction is the loss's maximum rather

than a collapse it can fall into; a variant adds the measurements back with a relative weight. *Graph regularization* keeps the published objective and adds the Dirichlet energy of (4) as a penalty. *The constrained functional* replaces the objective by the one (4) minimizes, with and without the published loss alongside. No training loss yields an adapter that can stand in for the solve: every one is beaten by applying the solve to the frozen prediction and training nothing (1.123), and the best adapter any of them produces reaches 1.146 against SWS's 1.088. With the solve retained, the differences shrink to a few percent and none reaches SWS. In every family the sweep selects the end of the grid at which the added term is weakest. The reason is representational: asked only to copy the converged field, with no measurements to satisfy, the adapter reaches 1.737 while the field itself is at 1.123. A rank-8 perturbation of the backbone's weights cannot express what a free field on the pixel grid can, so the two stages stay separate.

Table 20: **Training the adapter to replace the solve.** Median MAE relative to SWS + SRP, without and with the solve at read-out; each row is the best of its own sweep, and "seed sd" is over three seeds. 25 images, 5 per benchmark (20 from the calibration set and 5 from DDAD's evaluation split).

	adapter alone		+ SRP
What the adapter is asked to fit	MAE	seed sd	MAE
no adapter	3.201	0.000	1.123
the published objective (Seo et al., 2026)	1.174	0.026	1.031
the converged field, copied	1.737	0.016	1.097
that field and the measurements	1.146	0.030	1.029
the measurements and a graph penalty	1.199	0.019	1.027
the constrained functional	1.356	0.029	1.008
that and the published objective	1.196	0.031	1.011
SWS *(ours)*	1.088	0.000	1.000

A second-order energy. The intermediate distance is where the solve trails adaptation (App. F), and (6) says why. A textbook alternative to a parametric term is the second-order energy $\|\mathbf{L}\,\delta\|^2$, whose null space contains linear functions, so its minimizer extrapolates a trend instead of relaxing to a weighted average—the thin plate to (4)'s membrane. It has no constant either. At this budget, however, it does not converge: squaring the Laplacian squares its condition number, and after 1,200 conjugate-gradient iterations, an order of magnitude more work than the deployed solve, the field more than 64 pixels from any measurement is still exactly the frozen prediction. Its error is worse than (4)'s on every benchmark and improves monotonically with iterations, the signature of an unconverged solver rather than a wrong prior. A multigrid preconditioner for the biharmonic operator might change this; at this budget, the three-unknown regression of (7) is the form of extrapolation that converges.

Solving in log space. Solving (4) on $\log \hat{f}$ makes the correction multiplicative. It helps on DDAD and costs more elsewhere, for a net loss, and it is redundant once (7) is present: the $\log \hat{f}$ term already carries the global part of a multiplicative correction.

Confidence-modulated regularization. Weighting the data term of the penalized form (Sec. 4.1) by the backbone's confidence output selects zero modulation on every sample. The question does not arise for (4), which has no data term to weight.

The iteration count of the penalized form. In the penalized form the iteration count is a regularizer: truncating conjugate gradients restricts the correction to a Krylov subspace in which the well-conditioned components appear first (Hansen, 1998), so accuracy rises and then falls and some K has to be chosen (Tab. 21, Fig. 9). The optimum differs between benchmarks and between the adapted and frozen predictions, where the residual is larger and less local. Choosing K *per sample* from held-out measured pixels is worse than one fixed count, because the measured pixels share the bias the solve exists to correct: held-out sparse agreement is not an unbiased estimate of dense error. This argument concerns *estimating* dense error from measured pixels; (9) estimates nothing but solves a stationarity condition of the training objective, which is defined on the measured pixels

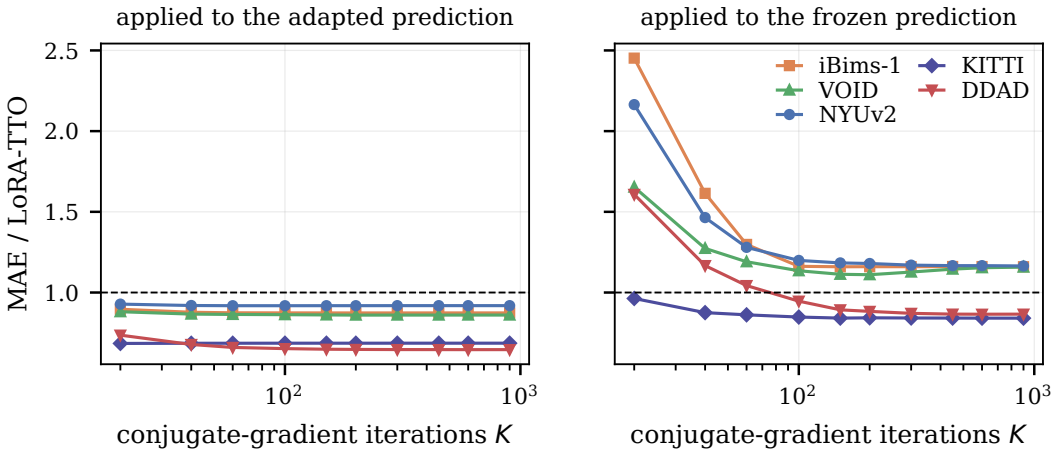

Figure 9: **Iteration count K of the *penalized* solve of Sec. 4.1, per benchmark.** Median MAE relative to LoRA-TTO against the conjugate-gradient iteration count, on an adapted and on the frozen prediction. 50 images.

by construction, as are all the gradients that follow it. None of this applies to (4), whose accuracy is monotone in the work spent (Tab. 11): its budget can be set from the latency available.

Table 21: **Iterations K of the *penalized* solve.** Median MAE relative to LoRA-TTO, over the core. **Bold and underlined** marks the value chosen on the calibration set. 50 images.

CG iterations K	20	40	60	100	200	450	900
applied to an adapted prediction	0.817	**0.774**	0.774	0.776	0.777	0.777	0.776
applied to the frozen prediction	1.862	1.291	1.154	1.082	**0.999**	0.970	0.955

I Qualitative Results

Fig. 10 shows what four configurations leave behind on one sample per benchmark, and Fig. 11 what they predict. The frozen backbone's error is regional rather than edge-shaped: whole surfaces sit at the wrong distance even after the scale-and-shift fit, which is what (2) says survives the alignment. After adaptation or the read-out, the error no longer fills regions; it collects along depth discontinuities and, on the driving rows, at the pixels farthest from any measurement, the group Tab. 12 isolates and (7) exists to reach. The read-out with no backward pass already produces a residual of the second kind, a visual counterpart of the claim that choosing among the minimizers is a different job from descending toward them. The quantitative evidence is Tab. 1 and Tab. 12.

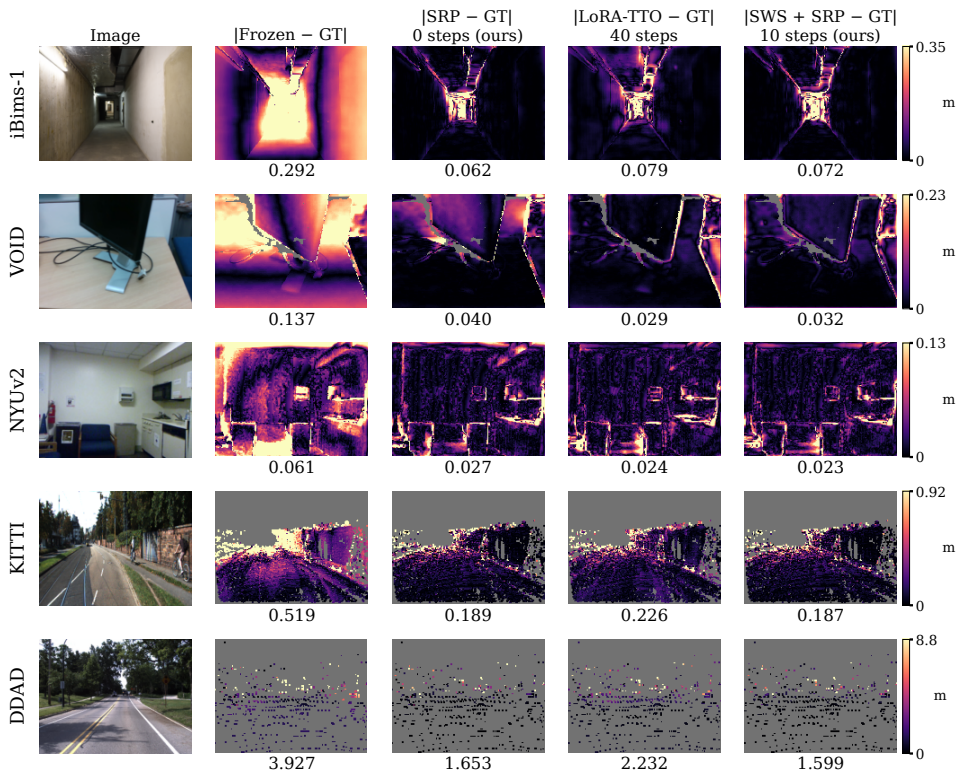

Figure 10: **Absolute error (m) on the samples of Fig. 11.** The scale (`magma`) is shared within a row and saturates at the 95th percentile of its four maps; error is defined only where ground truth exists (pixels without ground truth are gray), which on KITTI and DDAD is 19% and 1.0% of pixels.

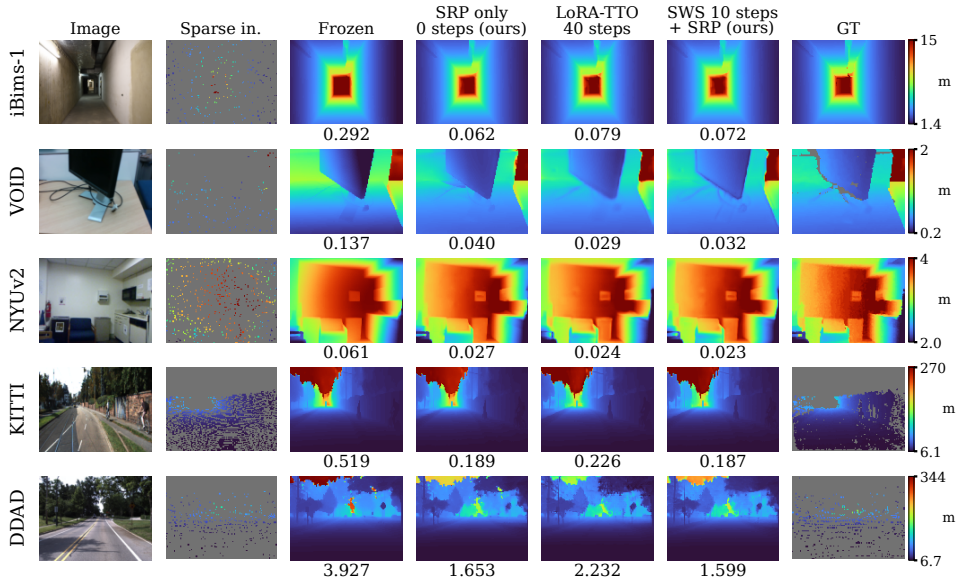

Figure 11: **Predicted depth on one sample per benchmark.** Four configurations, with the per-sample MAE (m) under each panel. The color scale (`turbo`) is fixed per row and shown in meters at the right; masked panels are dilated for visibility.

J LIMITATIONS

Intermediate distance. Without adaptation, SRP trails the 40-step run between 4 and 16 pixels from a measurement on every benchmark (Tab. 12), and this group holds most evaluated pixels on four of the five. A smooth interpolant cannot invent the structure missing there, which is what the 10 adaptation steps supply; an extrapolating field that converges at this budget (App. H.2) is the natural next step.

Feature-driven sparsity. VOID is the benchmark our full-budget configurations move least, the only one where the parametric term costs rather than helps, and the one where a richer basis does most harm. One property accounts for all three: its points come from a visual-inertial tracker and sit on trackable corners and textured structures, so the measured set concentrates where the affinity graph of (4) is weakly connected and is a biased sample of the scene, and anything fitted or selected on it generalizes least well. Affinities that stay informative under such patterns are the clearest open problem this work leaves.

The tail and the operating point. Without adaptation, SRP improves MAE on three of the five benchmarks but is behind on RMSE on all five, because a squared metric is priced by the tail the field reaches least well; the 10-step configuration is the first point on the curve that is ahead on both. At 10 steps, SWS ends worse than the frozen backbone on 1.5% of samples against the baseline's 0.5% (App. G).

Scope. The pipeline is not real-time: at the adaptation-free point it runs at 9.2 frames per second at VGA and 1.0 at FHD on one RTX 4090, and we state cost claims as ratios against the baseline. The analysis covers any aligned masked regression, but we instantiate it with low-rank adaptation on two backbones, and the diffusion-prior baselines are not run (App. C).