TUBULAR LOW-RANK LAPLACE APPROXIMATIONS FOR BAYESIAN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Laplace approximations provide post-hoc uncertainty estimates, but a single local Gaussian around a maximum-a-posteriori (MAP) solution can poorly represent curved, anisotropic low-loss regions. We introduce Tubular Low-Rank Laplace (TRL), a single-checkpoint post-hoc posterior approximation that combines a discrete low-loss spine with transverse Gaussian uncertainty across the full network. Starting from one MAP checkpoint, TRL uses stochastic cross-entropy Hessian-vector products to select a low-rank positive-curvature subspace as a Fisher/GGN-motivated proxy, transports the basis along a predictor–corrector spine, and samples validation-scaled transverse perturbations around uniformly selected anchors. Predictions use the resulting nonlinear networks, avoiding full curvature matrices and repeated local decompositions. On CIFAR-100, TRL improves NLL and calibration over the evaluated last-layer and rank-matched all-weight Laplace baselines and SWAG-Diag. Across five independently trained checkpoints, mean top-label ECE is 0.013, versus 0.096 for MAP and 0.028 after temperature scaling. On ImageNet/ResNet-50, the MAP-centered TRL tube reduces ECE relative to the deterministic checkpoint in all ten evaluation splits, with small NLL and accuracy costs; temperature scaling achieves lower NLL and ECE. Ablations show that transverse uncertainty accounts for most gains around deeply trained CIFAR-100 checkpoints. In small-data fine-tuning, the spine contains functionally distinct anchors; uniform mixing increases predictive diversity while retaining mean NLL and Brier comparable to a validation-selected single anchor, without an additional anchor-selection search.

1 INTRODUCTION

Laplace approximations are attractive post-hoc uncertainty methods for neural networks. Starting from a trained maximum a posteriori (MAP) solution, they fit a local Gaussian using curvature of the negative log posterior and can therefore be attached to an existing checkpoint without changing the original training procedure. Scalable variants use diagonal, Kronecker-factored, last-layer, or low-rank curvature approximations (Martens & Grosse, 2015; Ritter et al., 2018; Daxberger et al., 2021a). Their main limitation is locality: a single Euclidean Gaussian around one MAP point can be poorly aligned with the posterior geometry of an overparameterized network.

That locality is especially restrictive when low-loss sets are connected and strongly anisotropic. A local ellipsoid can represent uncertainty around one point, but it has no explicit way to allocate support along a curved low-loss direction. Linearized Laplace improves how local weight uncertainty is propagated to predictions (Immer et al., 2021b; Antorán et al., 2022; Daxberger et al., 2021a; Roy et al., 2024), yet it remains tied to a local approximation around the MAP. Geometric and Riemannian posterior methods address this structure more directly through Fisher/GGN metrics, geodesic transport, or learned low-loss manifolds (Amari, 1998; Bonnabel, 2013; Kristiadi et al., 2023; Léger & Vialard, 2024; Bergamin et al., 2023; Yu et al., 2024; Dold et al., 2025; Da Costa et al., 2025). Their more faithful forms, however, require operations that are difficult to scale, such as repeated local curvature fitting, geodesic construction, or MCMC over a connected manifold. This motivates a simpler question: can we preserve the useful longitudinal/transverse decomposition while keeping the inference procedure post hoc and low rank?

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.

We introduce *Tubular Low-Rank Laplace* (TRL), a posterior approximation built from two support components around one MAP checkpoint. Stochastic cross-entropy Hessian-vector products identify a low-rank, full-network transverse subspace through leading positive-curvature directions used as a Fisher/GGN-motivated proxy. A discrete predictor–corrector spine provides longitudinal support through a nearby low-loss corridor. TRL samples validation-scaled Gaussian perturbations in the transverse subspace around stored spine anchors and evaluates the resulting nonlinear networks. The construction avoids full curvature matrices, exact geodesics, and a fresh local decomposition at every anchor.

The experiments separate the roles of these two components. On CIFAR-100, TRL improves NLL and calibration over the evaluated last-layer and rank-matched all-weight Laplace baselines and SWAG-Diag; random rank-matched controls remain close to MAP. Around deeply trained CIFAR checkpoints, a MAP-centered transverse tube captures most of the gain. In small-data fine-tuning, the spine carries stronger functional variation: uniform mixing increases predictive diversity while retaining mean NLL and Brier comparable to a validation-selected single anchor without a separate anchor-selection search. Across ten ImageNet/ResNet-50 splits, the MAP-centered TRL tube reduces ECE in every split, while temperature scaling attains lower NLL and ECE; Deep Ensembles retain the best reported mean clean CIFAR-100 accuracy, NLL, and Brier.

This paper makes three contributions:

- We introduce TRL, a single-checkpoint post-hoc posterior approximation that combines a curvature-selected, full-network low-rank transverse sampler with a discrete low-loss spine and nonlinear posterior prediction.
- We isolate the contributions of transverse sampling, longitudinal mixing, and amortized basis transport using single-checkpoint/full-spine, random-support, fixed/transported/fresh-basis, and fine-tuning diagnostics. The results show different relative contributions across the evaluated regimes: transverse support accounts for most gains around deeply trained CIFAR checkpoints, whereas fine-tuning exhibits stronger functional variation along the spine.
- We evaluate TRL on toy problems (Appendix D), CIFAR-100, CIFAR-100-C, CIFAR-10H human-label alignment, architecture checks, and an ImageNet scale-check (Appendix J), including rank-matched all-weight low-rank Laplace, temperature scaling, random-subspace controls, and ablations of rank, spine extent, tube scale, curvature refresh, prior structure, and BatchNorm consistency.

Throughout, we distinguish this geometric idealization from the discrete, transported low-rank algorithm used in experiments. The supplementary material provides anonymized executable code, manuscript source, and compact provenance artifacts used to audit and reproduce the reported experiments.

2 Background

We consider a Bayesian neural network with prior $p(\theta)$ and likelihood $p(\mathcal{D} \mid \theta)$. Let $\mathcal{L} = -\log p(\mathcal{D} \mid \theta) - \log p(\theta)$ and $H = \nabla^2_\theta \mathcal{L}(\theta_{\mathrm{MAP}})$. ELA fits $\mathcal{N}(\theta_{\mathrm{MAP}}, H^{-1})$, whereas LLA uses $f_{\theta_{\mathrm{MAP}}+\delta}(x) \approx f_{\theta_{\mathrm{MAP}}}(x) + J_{\mathrm{MAP}}(x)\delta$, where δ is a parameter perturbation and $J_{\mathrm{MAP}}(x)$ is the output Jacobian at the MAP. With $g_\theta = \nabla_\theta \log p_\theta(y \mid x)$, the Fisher is $F(\theta) = \mathbb{E}[g_\theta g_\theta^\top]$. Fisher/GGN are PSD sensitivity operators; TRL instead retains the positive spectrum of stochastic cross-entropy Hessian HVPs, so they motivate rather than equal the computed operator. Appendix A gives details.

3 Geometric motivation

Deep-network loss landscapes contain highly anisotropic low-loss regions: many directions are near-invariances of the realized function, while others correspond to genuine mode-connectivity directions that can change predictions at comparable loss. TRL uses this distinction operationally. Because posterior density depends on both the likelihood and the prior, low training loss alone does not determine posterior mass. TRL uses the verified low-loss corridor as a support-allocation heuristic, while the amount of transverse spread within the selected support is controlled separately. With a fixed posterior-rank budget, locally flat symmetry directions can absorb substantial parameter variance

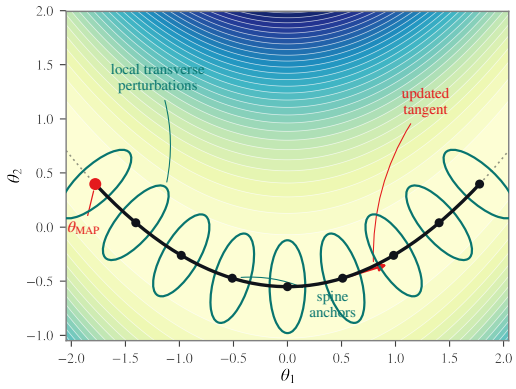


Figure 1: **Practical TRL support.** Starting from the MAP, TRL stores a discrete predictor–corrector spine and samples low-rank transverse Gaussian perturbations around its anchors. The tangent is updated from the realized spine displacement, while the transverse frame is transported and re-orthogonalized without a fresh local curvature decomposition at every anchor. Path curvature and fiber sizes are exaggerated for legibility.

while inducing little predictive diversity. TRL therefore allocates transverse rank to data-sensitive directions and represents low-loss longitudinal motion separately through the spine.

Figure 1 illustrates how TRL selects a low-rank set of data-sensitive transverse directions with positive-spectrum cross-entropy Hessian HVPs, used as a Fisher/GGN-motivated curvature proxy, constructs a discrete low-loss spine in the complementary region, and samples a calibrated low-rank posterior around the resulting support. The relative contribution of the two support axes changes with the training setting: the spine represents longitudinal support, and our diagnostics quantify when the traversed corridor carries functional variation. Appendix B gives the full geometric motivation, figures, Fisher-motivated tangent/normal decomposition, and continuous tubular idealization.

4 Tubular Low-Rank Laplace Approximation

TRL is a geometry-motivated, scalable post-hoc approximation to the weight-space posterior. The continuous tube is a conceptual reference; the implemented method is a discrete, transported low-rank approximation designed for scalable posterior predictive sampling from one MAP checkpoint.

4.1 Practical discrete TRL approximation

The implemented TRL posterior replaces the continuous tube by a discrete scalable approximation. It consists of:

1. a finite predictor–corrector spine $\{\gamma_t\}_{t=0}^{T-1}$ of T stored low-loss anchors;
2. a low-rank transverse basis initialized at the MAP and transported along the spine;
3. local Gaussian transverse perturbations applied at discrete spine anchors;
4. architecture-dependent batch-normalization recalibration, required when sampled networks contain BatchNorm layers.

Starting from one trained MAP, TRL computes dominant transverse structure once, then transports, projects, and re-orthogonalizes it along the spine. This amortization assumes that, over sufficiently short steps, the stiff subspace remains informative. Appendix A.4 details the predictor–corrector update, projection–QR transport, and discrete mixture.

Spine construction. Starting from $\gamma_0 = \theta_{\mathrm{MAP}}$, Lanczos/HVPs identify the stiff transverse subspace N_0, and the initial unit tangent satisfies $v_{\|,0} \in \mathrm{span}(N_0)^{\perp}$. For large-network CIFAR/ImageNet runs, $v_{\|,0}$ is a seeded random vector projected into this complement and normalized; toys instead use the smallest-eigenvalue eigenvector. The corrector subtracts the gradient component orthogonal to

the tangent; norm control limits the displacement, whose normalization gives the next tangent. For the large-network classification implementation, this gradient uses the mini-batch mean cross-entropy $\ell_{\mathcal{B}_t}(\theta) = -|\mathcal{B}_t|^{-1} \sum_{(x,y)\in\mathcal{B}_t} \log p_\theta(y \mid x)$, without a prior-gradient term. This differs from the negative log posterior $\mathcal{L}$; the prior enters the transverse precision separately. The spine is not an exact geodesic or smallest-eigenvector sequence; its nominal extent is $(T-1)\Delta s$.

Transported transverse structure. At each spine anchor γ_t, let $N_t \in \mathbb{R}^{K\times k_\perp}$ denote the low-rank transverse basis used for sampling, where K is the number of trainable parameters and $k_\perp$ is the transverse rank. The columns of N_t are the parameter-space directions in which TRL applies Gaussian transverse perturbations. At the MAP, N_0 is computed from dominant data-sensitive curvature directions using implicit curvature-vector products. These directions define the stiff subspace avoided by longitudinal motion; the prior affects scale, not selection. Along the spine, the basis is transported rather than recomputed. After each predictor–corrector update, we refresh the tangent from the realized displacement. Here d_t is the realized predictor–corrector displacement from γ_t to γ_{t+1}, whose predictor term follows $v_{\parallel,t}$; normalizing d_t gives the next tangent:

$$d_t = \gamma_{t+1} - \gamma_t, \qquad v_{\parallel,t+1} = \frac{d_t}{\|d_t\|}. \tag{1}$$

Using the refreshed tangent from Eq. (1), we project the previous transverse frame away from this tangent and re-orthogonalize:

$$\widetilde{N}_{t+1} = \left(I - v_{\parallel,t+1} v_{\parallel,t+1}^\top\right) N_t, \qquad N_{t+1} = \mathrm{qf}\left(\widetilde{N}_{t+1}\right). \tag{2}$$

Equation (2) defines the transported-frame update; $\mathrm{qf}(\cdot)$ denotes the orthonormal Q factor from a thin QR decomposition. The resulting N_t should therefore be interpreted as a transported sampling frame, not as the eigenbasis of a freshly decomposed local curvature matrix at γ_t.

The practical factor $L_{\perp,t}$ is diagonal in transported coordinates. If $\widehat{\lambda}_i$ is the ith retained MAP curvature value, $n_{t,i}$ is the ith column of N_t, and Λ_{prior} is the diagonal prior-precision matrix, the implementation uses

$$L_{\perp,t} = \mathrm{diag}_{i=1}^{k_\perp}\left[\max\left\{\widehat{\lambda}_i + n_{t,i}^\top \Lambda_{\mathrm{prior}} n_{t,i}, 10^{-6}\right\}^{-1/2}\right]. \tag{3}$$

In Eq. (3), the curvature values are fixed at the MAP, while the prior projection follows the transported frame. Here $\widehat{\lambda}_i$ are eigenvalues of the mini-batch-mean cross-entropy Hessian (per-example scale), whereas Λ_{prior} inherits a prior fitted by marginal likelihood from likelihood curvature accumulated over a 5,000-example subset; $\beta_\perp$ and c therefore calibrate a tempered transverse Gaussian rather than a matched full-data Laplace covariance. This is an inverse-square-root precision factor, not a dense Cholesky factor of freshly recomputed local curvature; $N_t L_{\perp,t} z_\perp$ maps a transverse Gaussian draw into parameter space.

Discrete longitudinal sampling. The practical sampler represents the longitudinal coordinate by sampling t uniformly from the stored anchors $\{0,\dots,T-1\}$ and then drawing a transverse Gaussian perturbation at that anchor; it does not sample a continuous $z_\parallel$ or evaluate the smooth longitudinal map of Appendix B.3. A practical sample has the form

$$\theta^{(r)} = \gamma_t + N_t\left(\beta_\perp L_{\perp,t} z_\perp^{(r)}\right), \quad z_\perp^{(r)} \sim \mathcal{N}(0, I_{k_\perp}), \tag{4}$$

where t indexes a spine anchor and $\beta_\perp$ is a global calibration/tempering multiplier on transverse spread. Thus longitudinal uncertainty is the discrete anchor distribution, while the Gaussian precision/covariance factor is applied only in the transported transverse coordinates; Appendix A.4 derives this continuous-to-discrete mixture explicitly. The implemented posterior is best understood as a discrete mixture of local low-rank Gaussian perturbations along a discrete spine, not as a smooth global Gaussian pushforward. The MAP-centered single-checkpoint tube is the collapsed-spine special case of the same family: the anchor distribution becomes a point mass at $\gamma_0 = \theta_{\mathrm{MAP}}$, while the curvature-selected transverse Gaussian is retained.

Empirical effect of amortization. Appendix H compares transported, fixed-MAP, single-checkpoint, and fresh-refresh variants. On CIFAR-100/ResNet-18, the stiff eigenspace rotates, but fresh refresh changes predictions only modestly and adds substantial cost; a MAP-centered tube already captures most gains. In this setting, computing the transverse structure once at the MAP provides the better cost–fidelity trade-off, although fresh local factors may matter in other regimes.

Batch-normalization recalibration. For architectures with batch normalization, each sampled network can have shifted activation statistics. Evaluating sampled weights with the MAP running statistics would mismatch the sampled parameters and BatchNorm state. We therefore use FixBN as an inference-time validity step: after sampling weights, batch-normalization statistics are recalibrated by forward passes on a training subset before evaluating predictions. FixBN is executed inside the posterior-prediction stages. Appendix C.6 explains the timing scopes and overlapping FixBN sub-timers used in Table 14, while Appendix G reports the rolling-mode batch-count sensitivity.

4.2 ALGORITHMIC SUMMARY

Algorithm 1 summarizes the large-network classification pipeline. The cross-entropy corrector constructs the spine, while MAP eigenvalues and projected prior precision define the transverse factors. Predictions average samples around one trained solution; no independently trained predictive models are introduced.

Algorithm 1 Practical TRL inference with transported transverse structure

Require: MAP θ_{MAP}, query input x_*, step Δs, stored-anchor count T, rank $k_\perp$, scale $\beta_\perp$, correction strength η_{corr}, sample count S.

Initialize: $\gamma_0 \leftarrow \theta_{\mathrm{MAP}}$.
Compute dominant transverse basis N_0 at the MAP via Lanczos/HVPs.
Draw a seeded random $v_{\parallel,0}$, project it into $\mathrm{span}(N_0)^\perp$, and normalize.
for $t = 0$ to $T-1$ **do**
    Form $L_{\perp,t}$ from MAP eigenvalues and the prior projected onto N_t.
    Store $\mathcal{T}_t = \{\gamma_t, N_t, L_{\perp,t}\}$.
    if $t < T-1$ **then**
        $d_t \leftarrow \Delta s\, v_{\parallel,t} - \eta_{\mathrm{corr}}(I - v_{\parallel,t} v_{\parallel,t}^\top)\nabla_\theta \ell_{\mathcal{B}_t}(\gamma_t)$.
        Apply norm control to d_t; set $\gamma_{t+1} \leftarrow \gamma_t + d_t$.
        Update $v_{\parallel,t+1}$ from the realized displacement; transport N_t by tangent projection and QR.
    end if
end for
for $r = 1$ to S **do**
    Sample t uniformly from $\{0, \ldots, T-1\}$.
    Sample transverse coordinate $z_\perp^{(r)} \sim \mathcal{N}(0, I_{k_\perp})$.
    $\theta^{(r)} \leftarrow \gamma_t + N_t(\beta_\perp L_{\perp,t} z_\perp^{(r)})$.
    if the network uses batch normalization **then**
        Recalibrate BN statistics for $\theta^{(r)}$ on a training subset.
    end if
    Compute prediction $y^{(r)} \leftarrow f(x_*; \theta^{(r)})$.
end for
return predictive Monte Carlo mean and posterior functional variance from $\{y^{(r)}\}_{r=1}^S$.

The sampler represents one longitudinal direction and $k_\perp$ transported transverse directions, not a complete posterior over all locally flat directions. It provides within-basin averaging and does not capture the across-basin diversity of Deep Ensembles.

4.3 PREDICTIVE MOMENTS AND INTERPRETATION

Given posterior samples $\theta^{(r)} \sim q_{\mathrm{TRL}}$, Eqs. (5)–(6) evaluate the nonlinear network directly, without linearizing it around the MAP. For classification, $f_\theta(x)$ denotes the post-softmax class-probability

vector. We report the Monte Carlo predictive mean and posterior functional variance:

$$\hat{\mu}_*(x_*) = \frac{1}{S}\sum_{r=1}^{S} f_{\theta^{(r)}}(x_*), \tag{5}$$

$$\widehat{\mathcal{V}}_f(x_*) = \frac{1}{S}\sum_{r=1}^{S} \|f_{\theta^{(r)}}(x_*) - \hat{\mu}_*(x_*)\|_2^2 . \tag{6}$$

For vector outputs, $\widehat{\mathcal{V}}_f$ is the trace of the empirical covariance of the predictive functions under the $1/S$ Monte Carlo normalization used in Eq. (6); for scalar regression it is the empirical posterior variance of the network output. It is the epistemic component of predictive variance, not the total $\mathrm{Var}(y_* \mid x_*, \mathcal{D})$, which also includes conditional observation or class uncertainty.

4.4 Scope and limitations of the approximation

The practical approximation makes several explicit choices. It uses a one-dimensional constructed spine, truncates transverse uncertainty to a low-rank subspace, transports transverse structure without a fresh local curvature factorization at every anchor, and uses a global scale parameter to calibrate transverse spread. Higher-dimensional low-loss manifolds are outside the present approximation. These choices are not incidental implementation details. They define the scalable approximation studied in this paper and should be evaluated as fidelity–cost trade-offs. A richer method could refresh curvature more frequently, use higher-dimensional spines, or combine within-basin geometry with cross-basin diversity, but such extensions would move the method to a different computational regime. Appendix M discusses these limitations and possible extensions in more detail.

5 Relation to existing methods

TRL is a post-hoc Laplace-family approximation, but it makes a different scalability cut from the Euclidean and linearized Laplace baselines used in our experiments. Here "Laplace-family" refers to the local Gaussian transverse fibers and curvature-informed construction; practical TRL is not a single second-order expansion of one globally specified posterior around an exact mode. Last-layer ELA/LLA restrict uncertainty to the classifier and factorize its curvature; TRL instead keeps full-network access through stochastic HVPs and restricts the posterior to a low-rank transverse subspace selected in the full parameter space, coupled to a discrete low-loss spine whose contribution varies across training settings.

TRL is also related to geometric and Riemannian posterior approximations, low-rank subspace and subnetwork inference, mode connectivity, loss-tunnel samplers, and geometric Gaussian approximations (Bergamin et al., 2023; Yu et al., 2024; Izmailov et al., 2020; Daxberger et al., 2021b; Garipov et al., 2018; Draxler et al., 2018; Dold et al., 2025; Da Costa et al., 2025; Faller & Martin, 2026). The computational distinction is that TRL amortizes support construction around one MAP solution and then samples closed-form low-rank Gaussian fibers, without claiming exact metric transport, repeated local covariance refits, or MCMC over a learned tunnel. Appendix O gives the detailed method-by-method positioning, including the different approximation target of Faller and Martin, the random/PCA/curve subspaces of Izmailov et al., and the endpoint assumptions of mode-connectivity methods.

To keep the comparison sets fixed, we distinguish four computational regimes. *Single-checkpoint post-hoc* methods start from one already-trained deterministic model and do not continue the original task training or fit additional independently optimized predictive models. Their post-hoc construction may still use training data, gradients, and nearby parameter-space updates; this computational regime includes MAP+TS, last-layer ELA/LLA, FM-LR-LLA ValScale, and TRL, with MAP serving as the deterministic reference. ELA/LLA, FM-LR-LLA, and TRL return posterior approximations, whereas MAP+TS is a scalar non-posterior calibration control. SWAG-Diag is *checkpoint-initialized* but follows a short additional optimization trajectory to collect diagonal weight statistics. MC Dropout is a *single-training, training-time stochastic* method whose predictive dispersion is introduced during training. Deep Ensembles require multiple independent trainings. Thus, "single-training" is broader than "single-checkpoint post-hoc," and TRL specifically asks how much posterior averaging can be recovered from one deterministic checkpoint by selecting a useful full-network low-rank support.

Table 1: CIFAR-100 clean test performance (five independently trained MAP checkpoints; mean ± sample SD; one TRL posterior per checkpoint). ECE is 15-bin equal-width top-label ECE. MAP+TS fits one positive temperature per checkpoint by clean-validation NLL and preserves the checkpoint argmax. FM-LR-LLA ValScale is the rank-matched all-weight Kronecker comparison; Appendix E.1 gives protocol details. Text comparisons use unrounded means; see Appendix A.5.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑
Deep Ensemble	0.779 ± 0.001	0.805 ± 0.007	0.017 ± 0.002	0.308 ± 0.003	0.862 ± 0.008
MAP	0.743 ± 0.003	1.040 ± 0.008	0.096 ± 0.002	0.370 ± 0.003	0.849 ± 0.021
MAP+TS	0.743 ± 0.003	0.960 ± 0.006	0.028 ± 0.004	0.355 ± 0.003	0.870 ± 0.022
SWAG-Diag	0.744 ± 0.003	1.040 ± 0.007	0.096 ± 0.003	0.369 ± 0.003	0.852 ± 0.020
MC Dropout	0.742 ± 0.007	0.953 ± 0.021	0.048 ± 0.005	0.357 ± 0.007	0.780 ± 0.027
ELA (last-layer)	0.724 ± 0.002	1.356 ± 0.013	0.262 ± 0.008	0.468 ± 0.004	0.874 ± 0.021
LLA (last-layer)	0.739 ± 0.002	1.492 ± 0.017	0.368 ± 0.008	0.521 ± 0.006	0.903 ± 0.008
FM-LR-LLA ValScale	0.743 ± 0.002	0.985 ± 0.006	0.041 ± 0.010	0.358 ± 0.003	0.843 ± 0.026
TRL (ours)	0.744 ± 0.003	0.951 ± 0.006	0.013 ± 0.002	0.353 ± 0.003	0.867 ± 0.021

6 EXPERIMENTS

Protocol. We compare TRL with MAP, last-layer ELA/LLA, FM-LR-LLA ValScale (Faller & Martin, 2026), SWAG-Diag, MC Dropout, and Deep Ensembles. TRL, MAP+TS, and FM-LR-LLA ValScale use held-out clean-validation NLL for scale/temperature selection; headline ELA/LLA instead use marginal-likelihood prior precision, while a validation-NLL prior grid sends both to $\lambda = 10^4$, effectively MAP (Appendix E.2). OOD AUROC uses predictive entropy. The main CIFAR-100 TRL setting is $k_\perp = 30$, $T = 40$, $\beta_\perp = 4.0$, $S = 25$, with 25-batch FixBN. Appendix C gives the complete split, seed, tuning, and baseline protocols, including the rolling-versus-reset FixBN sensitivity audit.

Experimental roles. Headline tables aggregate independently trained checkpoints; fixed-MAP ablations vary posterior draws, while single-checkpoint/fresh-refresh and FixBN/replay studies are mechanism or provenance diagnostics rather than benchmark replicates.

At CIFAR scale, ELA/LLA use last-layer K-FAC/KRON curvature; FM-LR-LLA ValScale is a checkpoint-, split-, and rank-matched matrix-free all-weight Kronecker reference with a validation-scaled rank-30 predictive-covariance subspace. The prior families differ: FM uses one scalar precision, whereas TRL uses its validation-selected two-block head/backbone prior. Appendix E.1 gives the FM diagnostics; Appendices K and H give random-support and transport controls.

6.1 CIFAR-100 CLEAN TEST PERFORMANCE

Table 1 reports the five-checkpoint clean comparison. Deep Ensembles retain the best reported mean accuracy, NLL, and Brier at five-training cost; TRL has the lowest reported mean clean 15-bin equal-width top-label ECE. MAP+TS preserves the checkpoint argmax and improves NLL/ECE/Brier to $0.960/0.028/0.355$; TRL remains lower at $0.951/0.013/0.353$. All five checkpoint-matched ECE differences favor TRL (mean difference -0.0147; TRL mean about 48% of MAP+TS). OOD ranking is distinct: last-layer LLA has the highest SVHN AUROC despite poor clean likelihood and calibration.

The matched low-rank controls narrow the source of the gain. FM-LR-LLA improves over MAP but remains above TRL in NLL, ECE, and Brier, while random rank-30 full-network controls remain close to MAP (Appendix K). Full-network access and low rank alone are therefore insufficient. The remaining differences include curvature-selected support, TRL's two-block prior, and nonlinear prediction. FixBN is the BatchNorm-consistency step required to evaluate sampled networks and does not define posterior support. MAP+TS shows that scalar recalibration explains most of the NLL/Brier correction, while TRL retains lower top-label ECE and sample-based functional diagnostics. Appendix E.2 confirms the top-label ordering under equal-width, equal-mass, and KS checks; classwise ECE slightly favors MAP+TS, and Deep Ensembles are numerically best under some estimators.

Table 2: CIFAR-100-C over 19 corruptions and five severities (mean ± sample SD over three seed-matched checkpoints). TRL uses the matched MAP checkpoint; MAP+TS fits one clean-validation temperature and freezes it for all corruptions.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓
Deep Ensemble	0.509 ± 0.001	2.292 ± 0.012	0.114 ± 0.002	0.650 ± 0.003
MAP	0.471 ± 0.004	2.776 ± 0.033	0.234 ± 0.003	0.747 ± 0.005
MAP+TS	0.471 ± 0.004	2.390 ± 0.041	0.118 ± 0.008	0.691 ± 0.007
SWAG-Diag	0.473 ± 0.005	2.774 ± 0.062	0.234 ± 0.006	0.745 ± 0.008
MC Dropout	0.475 ± 0.006	2.554 ± 0.060	0.172 ± 0.008	0.707 ± 0.010
TRL (ours)	0.473 ± 0.004	2.464 ± 0.047	0.118 ± 0.006	0.689 ± 0.005

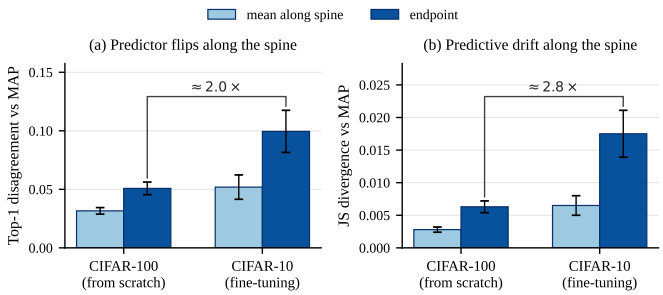


Figure 2: **Longitudinal variation differs across regimes.** Endpoint top-1 disagreement and JS drift are about $2\times$ and $2.8\times$ larger under small-data fine-tuning (sample SD over five and ten runs).

6.2 CIFAR-100-C ROBUSTNESS AND CALIBRATION

Table 2 averages 19 corruptions and five severities over three independently trained, seed-matched checkpoints. Deep Ensembles lead every reported mean. Among posterior approximations, TRL has the lowest NLL, ECE, and Brier at MAP-like accuracy. The frozen clean-validation MAP+TS control has lower NLL than TRL in all three paired checkpoints, essentially the same aggregate ECE, and a slightly higher Brier; neither method is tuned on corruption. The comparison therefore separates two effects: scalar recalibration remains a strong control under shift, while TRL additionally provides a sampled full-network posterior and support diagnostics.

6.3 ABLATIONS AND REGIME DIAGNOSTICS

ECE varies non-monotonically with transverse rank (Table 17), with $k_\perp = 30$ giving the best ECE among the five evaluated ranks. Transported and fixed-MAP bases are nearly tied, fresh refresh changes little, and a single-checkpoint tube captures most gains in the main CIFAR regime.

Figure 2 shows greater longitudinal variation under small-data adaptation. Endpoint top-1 disagreement/JS are $5.08\% \pm 0.54\%/0.0063 \pm 0.0009$ on CIFAR-100 from scratch and $9.95\% \pm 1.80\%/0.0175 \pm 0.0036$ after CIFAR-100→CIFAR-10 fine-tuning. In fine-tuning, full-spine mixing increases functional diversity while retaining mean NLL and Brier comparable to the validation-selected single-anchor control. The mixture keeps support across the discovered corridor without a separate anchor-selection rule. Appendices H and I give the full comparisons; Appendices G and J give the remaining ablations and ImageNet checks.

Two further controls separate support choice from local curvature tracking. Along the CIFAR-100 spine, transported–fresh top-30 eigenspace overlap falls to 0.198 ± 0.021 at the endpoint, with a mean principal angle of $70.8^\circ \pm 1.8^\circ$; the fixed-MAP basis is essentially no closer to the fresh local eigenspace. Despite that rotation, fixed and transported bases give nearly identical aggregate predictive metrics. Recomputing the eigenspace at every anchor changes NLL only modestly and adds 89.2 ± 7.7 minutes of offline construction per seed (Appendix H). In this benchmark, the useful

object is therefore the MAP-identified support itself, not accurate tracking of the local stiff eigenspace along the whole path.

Table 3: Compact regime comparison. CIFAR-100 rows summarize Tables 25 and 26 over three checkpoint-specific runs; fine-tuning rows reproduce Table 27 over ten seeds. Comparisons should be made within each block.

Regime	Variant	NLL ↓	ECE ↓	Brier ↓	Avg. func. var.
CIFAR-100	single checkpoint	0.958 ± 0.005	0.015 ± 0.003	0.357 ± 0.002	–
CIFAR-100	transported full spine	0.956 ± 0.004	0.016 ± 0.003	0.355 ± 0.001	–
CIFAR-100	fresh refresh	0.955 ± 0.008	0.013 ± 0.003	0.356 ± 0.002	–
Fine-tuning	single checkpoint	0.649 ± 0.014	0.019 ± 0.003	0.315 ± 0.005	0.00078 ± 0.00010
Fine-tuning	best single (val.)	0.647 ± 0.015	0.019 ± 0.004	0.314 ± 0.006	0.00094 ± 0.00008
Fine-tuning	endpoint single	0.647 ± 0.014	0.019 ± 0.004	0.314 ± 0.006	0.00096 ± 0.00011
Fine-tuning	full spine	0.647 ± 0.015	0.017 ± 0.003	0.314 ± 0.006	0.00230 ± 0.00043

As Table 3 shows, in the fine-tuning diagnostic center relocation accounts for most of the mean NLL/Brier change: both the validation-selected and endpoint-centered tubes are close to the full-spine mixture. Mixing primarily increases functional diversity, raising average functional variance to 2.30×10^{-3} from roughly 0.9×10^{-3} for the single-anchor controls.

The endpoint is a strong ex-post control, but its location is fixed by path length and step schedule rather than by a predictive criterion. Validation selection provides such a criterion by evaluating six candidate anchors; uniform mixing instead retains support across the discovered corridor without selecting one anchor. The larger fine-tuning disagreement also coincides with more loss drift (endpoint validation cross-entropy change 0.0594 ± 0.0239, versus about 0.010 from scratch), so functional variation and low loss should be read separately.

Random rank-30 bases remain close to MAP across a $20\times$ scale range under the same validation-scale selection and FixBN protocol, while curvature-selected TRL has an interior operating range and degrades when over-scaled (Appendix K). The rank sweep points in the same direction: support choice and scale both matter; full-network access and low rank alone do not reproduce the effect.

Across the two evaluated regimes, the balance between transverse and longitudinal support changes. A MAP-centered tube is sufficient when longitudinal variation is weak, whereas the full spine preserves additional diversity when the corridor contains distinct predictors of comparable quality. The ordering is regime-specific: representative counterexamples are LLA on OOD AUROC; MAP+TS on CIFAR-100-C NLL and ImageNet NLL/ECE; MC Dropout on WideResNet-16-4 accuracy/NLL/Brier and VGG-11-BN accuracy/NLL/ECE/Brier; and Deep Ensembles on clean CIFAR-100 accuracy/NLL/Brier and ambiguous-CIFAR-10H soft Brier/JS (Appendix F.1).

7 Conclusion

TRL combines a curvature-selected full-network transverse subspace with a discrete low-loss spine in a single-checkpoint posterior approximation. On CIFAR-100, it improves NLL and top-label calibration over the evaluated Laplace baselines, with most gains already captured by the MAP-centered tube. On ImageNet, that tube lowers checkpoint ECE, but temperature scaling is lower in NLL and ECE. The longitudinal contribution is regime-dependent: in small-data fine-tuning, spine mixing increases functional diversity while retaining mean NLL and Brier comparable to the evaluated single-anchor controls. Together, these diagnostics separate transverse sampling, center relocation, and longitudinal mixing while keeping the geometric assumptions empirically testable.

Reproducibility Statement

Appendices C, E, and G document protocols and provenance; Appendix F.1 covers CIFAR-10H. The supplement provides anonymized code, source, checksums, artifacts, and reproduction commands. No test, corruption, or CIFAR-10H labels are used for selection; Appendix C documents the rolling-FixBN state dependency and replay audits.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Javier Antorán, David Janz, James U Allingham, Erik Daxberger, Riccardo Barbano, Eric Nalisnick, and José Miguel Hernández-Lobato. Adapting the linearised laplace model evidence for modern deep learning. In *International Conference on Machine Learning*, pp. 796–821. PMLR, 2022.

Federico Bergamin, Pablo Moreno-Muñoz, Søren Hauberg, and Georgios Arvanitidis. Riemannian Laplace approximations for Bayesian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Silvère Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Tristan Cinquin, Marvin Pförtner, Vincent Fortuin, Philipp Hennig, and Robert Bamler. FSP-Laplace: Function-space priors for the Laplace approximation in Bayesian deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Nathaël Da Costa, Bálint Mucsányi, and Philipp Hennig. Geometric Gaussian approximations of probability distributions. *arXiv preprint arXiv:2507.00616*, 2025.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Erik Daxberger, Eric Nalisnick, James U. Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2510–2521, 2021b.

Daniel Dold, Julius Kobialka, Nicolai Palm, Emanuel Sommer, David Rügamer, and Oliver Dürr. Paths and ambient spaces in neural loss landscapes. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 10–18, 2025.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1308–1317, 2018.

Josua Faller and Jörg Martin. Low rank based subspace inference for the laplace approximation of bayesian neural networks. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL `https://openreview.net/forum?id=RTwTXQX4gq`.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1050–1059, 2016.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Mohammad Emtiyaz Khan. Scalable marginal likelihood estimation for model selection in deep learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021a.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 703–711, 2021b.

Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 1169–1179, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Agustinus Kristiadi, Felix Dangel, and Philipp Hennig. The geometry of neural nets' parameter spaces under reparametrization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Flavien Léger and François-Xavier Vialard. A geometric Laplace method. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 226–234, 2024.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9617–9626, 2019.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Hrittik Roy, Marco Miani, Carl Henrik Ek, Philipp Hennig, Marvin Pförtner, Lukas Tatzel, and Søren Hauberg. Reparameterization invariance in approximate Bayesian inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Levent Sagun, Utku Evci, V. Uğur Güney, Yann N. Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2018. ICLR 2018 Workshop.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Lukas Nicola Tatzel, Bálint Mucsányi, Osane Hackel, and Philipp Hennig. Debiasing mini-batch quadratics for applications in deep learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Hanlin Yu, Marcelo Hartmann, Bernardo Williams Moreno Sanchez, Mark Girolami, and Arto Klami. Riemannian Laplace approximation with the Fisher metric. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 820–828, 2024.

A Background on Laplace, Fisher/GGN, and Calibration Metrics

This appendix provides a short self-contained background for readers less familiar with Bayesian neural networks, Laplace approximations, and the geometric language used in the main text. The goal is not to survey the full literature, but to define the objects that appear in TRL and in the experimental baselines.

A.1 Bayesian prediction and MAP checkpoints

A Bayesian neural network places a prior $p(\theta)$ over network parameters and predicts by marginalizing over the posterior (Gal, 2016; Kendall & Gal, 2017; Rasmussen & Williams, 2006),

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid \mathcal{D})\, d\theta.$$

For modern deep networks this integral is intractable. Practical Bayesian deep learning therefore replaces $p(\theta \mid \mathcal{D})$ by a tractable approximation built around a trained solution. The maximum a posteriori (MAP) parameter θ_{MAP} minimizes the negative log posterior,

$$\mathcal{L}(\theta) = -\log p(\mathcal{D} \mid \theta) - \log p(\theta),$$

and serves as the deterministic checkpoint to which post-hoc uncertainty methods attach posterior mass. In the experiments, MAP denotes this deterministic model and also the center from which Laplace, SWAG-Diag, and TRL variants start.

A useful distinction is between *parameter uncertainty* and *predictive uncertainty*. A posterior over weights is only useful through the distribution over functions it induces. Two parameter perturbations can be large in Euclidean norm yet nearly identical functionally because of symmetries, normalization effects, or flat directions; conversely, a small movement in a data-sensitive direction can change predictions substantially. This is why the paper repeatedly separates geometric movement in parameter space from functional diagnostics such as disagreement, Jensen–Shannon divergence, and functional variance.

A.2 Euclidean and linearized Laplace

The Euclidean Laplace approximation (ELA) fits a Gaussian around the MAP by using the local quadratic approximation

$$\mathcal{L}(\theta_{\mathrm{MAP}} + \delta) \approx \mathcal{L}(\theta_{\mathrm{MAP}}) + \tfrac{1}{2}\delta^\top G_0 \delta, \qquad q_{\mathrm{ELA}}(\theta) = \mathcal{N}(\theta_{\mathrm{MAP}}, G_0^{-1}). \tag{7}$$

In Eq. (7), $\delta \in \mathbb{R}^K$ is a parameter displacement and G_0 is a curvature matrix such as a Hessian, Fisher, or generalized Gauss–Newton (GGN) matrix plus prior precision (MacKay, 1992; Daxberger et al., 2021a). If u_i is an eigenvector of the likelihood-curvature component with eigenvalue λ_i and an isotropic prior contributes precision λ, the Laplace variance along u_i is approximately $(\lambda_i + \lambda)^{-1}$. Thus high-curvature directions receive small posterior variance, while low-curvature directions receive larger variance. In a neural network, however, the full $K \times K$ curvature is too large to materialize at CIFAR scale, so practical Laplace baselines use restrictions or approximations such as last-layer curvature, Kronecker factorizations, diagonal structure, or low-rank subspaces.

For classification, let $a_\theta(x)$ denote the pre-link logit vector and $p_\theta(x)$ the post-link class-probability vector. Linearized Laplace approximation (LLA) uses the same local Gaussian weight posterior but propagates it through the first-order Taylor expansion of the logits,

$$a_{\theta_{\mathrm{MAP}}+\delta}(x) \approx a_{\theta_{\mathrm{MAP}}}(x) + J^a_{\theta_{\mathrm{MAP}}}(x)\delta. \tag{8}$$

Here $J^a_{\theta_{\mathrm{MAP}}}(x) = \partial a_\theta(x)/\partial\theta|_{\theta=\theta_{\mathrm{MAP}}}$ is the logit Jacobian at the MAP. For regression, the analogous expansion applies directly to the model output. Equation (8) therefore yields a Gaussian approximation in logit space for classification, or function space for regression, before the likelihood/link approximation is applied (Kristiadi et al., 2020; Daxberger et al., 2021a). Thus ELA and LLA differ mainly in how predictions are computed from the local Gaussian: ELA evaluates sampled nonlinear networks, whereas LLA evaluates the linearized model, often with a probit or moment-matching approximation for classification. Section 4.3, by contrast, uses $f_\theta(x) = p_\theta(x)$ for classification when defining TRL predictive moments. In the toy experiments, full-network curvature is small enough that ELA and LLA can be run as full-parameter baselines. In the CIFAR-scale experiments, the reported ELA/LLA baselines use last-layer Kronecker-factored Laplace for tractability, as specified in Appendix C.

A.3 FISHER/GGN GEOMETRY AND DATA-SENSITIVE DIRECTIONS

Information geometry equips parameter space with a local metric derived from predictive sensitivity rather than Euclidean distance (Amari, 1998; Bonnabel, 2013; Kristiadi et al., 2023). In the Fisher/GGN view, the quadratic form

$$v^\top F(\theta) v$$

measures how strongly a small parameter displacement v changes the model's predictions on the data, at least locally. Large values indicate data-sensitive or stiff directions; small values indicate directions that are approximately functionally invariant under the local metric. For a generalized Gauss–Newton matrix, this statement is especially transparent because F has the form of a Jacobian-transpose–curvature–Jacobian operator, so $Fv \approx 0$ is associated with $Jv \approx 0$ in the relevant data-weighted sense.

TRL uses this sensitivity picture to motivate its tangent/transverse decomposition. A low-loss region may contain directions along which the network can move while leaving training loss and predictions nearly unchanged, together with transverse directions where predictions change rapidly. The continuous tube in the main text idealizes these roles as a longitudinal spine with transverse Gaussian uncertainty. The implementation uses a lower-cost approximation: Hessian-vector products select a low-rank set of dominant data-sensitive directions, a finite spine is constructed in a low-curvature complement, and the transverse representation is transported along that spine. The geometric language organizes posterior support into components that can be approximated at scale; it is not a claim of exact geodesic sampling.

A.4 CONTINUOUS TRL AND ITS DISCRETE APPROXIMATION

The continuous tubular idealization writes the TRL posterior as a longitudinal mixture of transverse Gaussian fibers,

$$q_{\mathrm{TRL}}(d\theta) = \int \left(\exp_{\gamma(s)}\right)_{\#} \mathcal{N}_\perp(0, \Sigma_\perp(s))\ \rho(ds). \tag{9}$$

The objects in Eq. (9) are as follows:

- $\gamma(s)$ is the point on the ideal low-loss spine at longitudinal coordinate s.
- $\mathcal{N}_\perp(0, \Sigma_\perp(s))$ is a Gaussian supported only in the normal/transverse space at that point.
- For a measurable map ϕ and measure μ, $\phi_{\#}\mu$ denotes the pushforward measure, $(\phi_{\#}\mu)(A) = \mu(\phi^{-1}(A))$; thus $(\exp_{\gamma(s)})_{\#}$ maps the transverse Gaussian back into parameter space.
- $\rho(ds)$ is the longitudinal mixing measure, i.e., the probability weight assigned to choosing position s along the spine.

Equation (9) is a geometric reference object: it separates longitudinal support from transverse uncertainty, but the implementation does not evaluate it by continuous quadrature. In Eq. (17) (Appendix B.3), the tubular map is the flat, first-order normal-chart approximation to the exponential-map idealization. Under a locally Euclidean chart, the pushforward by $\exp_{\gamma(s)}$ is approximated by the affine map $\gamma(s) + N(s)L_\perp(s)z_\perp$. The released implementation uses this affine chart at each spine anchor, written below as ψ_t.

The implemented sampler replaces this object by an empirical mixture over discrete spine anchors. With discrete spine anchors $\{\gamma_t\}_{t=0}^{T-1}$ and transported frames N_t, the longitudinal measure is approximated by the uniform empirical measure,

$$\rho(ds) \approx \frac{1}{T} \sum_{t=0}^{T-1} \delta_{s_t}(ds), \tag{10}$$

where s_t is the longitudinal coordinate associated with anchor γ_t in the continuous idealization and δ_{s_t} is the Dirac probability measure concentrated at s_t. Substituting the empirical measure in Eq. (10) gives the practical posterior

$$q^{(T)}_{\mathrm{TRL}}(d\theta) = \frac{1}{T} \sum_{t=0}^{T-1} (\psi_t)_{\#} \mathcal{N}(0, I_{k_\perp}), \qquad \psi_t(\xi) = \gamma_t + N_t(\beta_\perp L_{\perp,t}\xi)\,. \tag{11}$$

In Eq. (11), $\xi \sim \mathcal{N}(0, I_{k_\perp})$ is the transverse coordinate and $L_{\perp,t}$ is the diagonal inverse-square-root precision factor of Eq. (3). The equation is the implemented discrete mixture: each posterior draw first samples an anchor index t uniformly, then draws a rank-$k_\perp$ transverse Gaussian perturbation in the transported coordinates. Uniform anchor weighting is a practical heuristic; it is not claimed to represent equal local evidence, equal posterior volume, or an exact quadrature rule along the corridor. The main CIFAR-100 configuration uses $T = 40$ discrete spine anchors to define this discrete mixture and $S = 25$ Monte Carlo draws from $q^{(T)}_{\mathrm{TRL}}$ to estimate posterior predictive probabilities.

The large-network classification spine uses mini-batch mean cross-entropy $\ell_{\mathcal{B}_t}$, defined in Section 4.1, rather than the negative log posterior $\mathcal{L}$. No prior gradient is added to this corrector; prior precision enters the transverse factors. If $v_{\parallel,t}$ is the current unit tangent, the un-clipped update has the form

$$\gamma_{t+1} \leftarrow \gamma_t + \Delta s\, v_{\parallel,t} - \eta_{\mathrm{corr}} \left(I - v_{\parallel,t} v_{\parallel,t}^\top\right) \nabla_\theta \ell_{\mathcal{B}_t}(\gamma_t). \tag{12}$$

Equation (12) is implemented with additional norm control in the released code. Here Δs is the prescribed longitudinal step size along the current tangent direction, and η_{corr} controls the strength of the projected gradient correction. The first term advances along the current longitudinal direction, while the projected corrector subtracts the component of the cross-entropy gradient orthogonal to the tangent in order to keep the trajectory in a low-loss corridor. This is not an exact geodesic update; it is a practical low-curvature-complement trajectory constructed from stochastic gradients and HVP-selected transverse directions.

After the realized update, the tangent and transported frame are refreshed using the actual displacement,

$$d_t = \gamma_{t+1} - \gamma_t, \qquad v_{\parallel,t+1} = \frac{d_t}{\|d_t\|}, \tag{13}$$

and the transverse basis is updated by projection and QR re-orthogonalization,

$$\widetilde{N}_{t+1} = \left(I - v_{\parallel,t+1} v_{\parallel,t+1}^\top\right) N_t, \qquad N_{t+1} = \mathrm{qf}\left(\widetilde{N}_{t+1}\right). \tag{14}$$

Here $\mathrm{qf}(A)$ denotes the orthonormal Q factor of a thin QR decomposition. Equations (13) and (14) define the projection–QR step that preserves orthogonality to the refreshed tangent, but it does not recompute the local Fisher/GGN eigenspace at γ_{t+1}. The diagnostics in Appendix H explicitly measure the consequences of this amortization by comparing transported, fixed-MAP, and freshly recomputed variants.

Finally, the posterior predictive distribution is estimated by Monte Carlo averaging,

$$\widehat{p}(y \mid x) = \frac{1}{S} \sum_{r=1}^{S} p(y \mid x, \theta^{(r)}), \qquad \theta^{(r)} \sim q^{(T)}_{\mathrm{TRL}}. \tag{15}$$

Equation (15) is the Monte Carlo estimator used after discretization. Thus the passage from the continuous object in Eq. (9) to the released TRL sampler involves three explicit approximations: the longitudinal mixing measure $\rho(ds)$ is replaced by the empirical uniform measure over the T discrete spine anchors; the normal exponential map is replaced by the local affine chart ψ_t; and the position-dependent transverse covariance is represented by transported frames and amortized diagonal inverse-square-root precision factors rather than freshly recomputed dense local Cholesky factors. The scalar $\beta_\perp$ enters at this discrete sampling stage as a validation-selected transverse temperature, not as an additional invariant of the continuous idealization. Finally, posterior prediction is estimated with S Monte Carlo draws from this discrete tube mixture. T controls the number of stored anchors and, at fixed Δs, the nominal longitudinal extent; S controls the Monte Carlo accuracy of the predictive estimate.

A.5 Calibration and probabilistic metrics

The experiments report accuracy, negative log likelihood (NLL), expected calibration error (ECE), Brier score, and, for OOD detection, predictive-entropy AUROC. Accuracy measures only the most likely class. NLL evaluates the probability assigned to the true label and strongly penalizes confident mistakes. The Brier score is the mean squared error between the predictive probability vector and the one-hot label, so it measures both calibration and sharpness of the predictive distribution.

ECE summarizes whether top-label confidence agrees with empirical top-label accuracy. Unless a table states otherwise, we use 15 equal-width confidence bins; within each bin, top-label ECE compares average maximum-class probability with average correctness and weights the absolute gap by bin frequency. ECE is easy to interpret but depends on binning and on whether calibration is summarized top-label-wise or classwise, so it should be read together with NLL and Brier score. Appendix E.2 reports equal-width, equal-mass, classwise, and Kolmogorov–Smirnov sensitivity checks and limits the headline claim to top-label calibration. OOD AUROC is different: it measures how well predictive entropy ranks in-distribution and out-of-distribution examples. A method can have strong OOD ranking but poor in-distribution calibration, or vice versa, which is why the tables report both probabilistic metrics and predictive-entropy AUROC when applicable. Unless explicitly marked as diagnostic experiments, tables report mean $\pm$ standard deviation over independent seeds. Tables generally use three decimal places; additional digits are retained where a study's reported comparisons require them, for example to prevent a nonzero standard deviation or a diagnostically relevant small difference from rounding to zero. Unless a table caption states otherwise, boldface marks the best unrounded mean among the rows listed in that table; values that tie after display rounding need not all be bold. Some tables intentionally use no boldface; absence of the mark is not a claim of ties. Numerical differences, percentages, and ratios stated in prose are likewise computed from the unrounded aggregates unless explicitly identified as display-level arithmetic. Where a table omits a method class, or where the surrounding text makes a narrower claim, the mark carries no scope beyond the rows shown.

B Geometric Motivation and Continuous TRL Idealization

In deep networks, the geometry of $\mathcal{L}(\theta)$ is characterised by two key features:

1. **Low-loss valleys.** Deep networks can exhibit several sources of weakly constrained motion. Continuous functional symmetries, such as positive-homogeneous rescalings, can generate locally flat tangent directions along which the represented function changes little. Discrete symmetries such as neuron permutations instead produce functionally equivalent parameter configurations but do not, by themselves, define a continuous tangent direction between them (Kristiadi et al., 2023). Approximate mode connectivity is distinct: low-loss paths can connect independently trained solutions that are functionally close but not identical, so motion along them can correspond to genuinely different predictors of comparable loss (Garipov et al., 2018; Draxler et al., 2018). A low-loss valley can mix continuous near-invariances with connectivity directions, with relative importance depending on architecture and dataset.
2. **Extreme anisotropy of curvature.** The loss is extremely flat along approximate invariance and connectivity directions and steep along data-sensitive directions. The Hessian spectrum thus exhibits a wide range of eigenvalues, with a bulk of near-zero values (Sagun et al., 2018; Maddox et al., 2019).

This distinction matters for what a longitudinal posterior construction can be expected to contribute. TRL does not assume that every longitudinal direction produces predictive diversity. It builds a discrete predictor–corrector spine to cover the low-loss corridor around the MAP. Appendix H verifies empirically that the stored anchor centers remain in a narrow low-loss region. Transverse Gaussian perturbations are then sampled around those anchors using a curvature-selected subspace; the low-loss anchor diagnostic does not certify every posterior draw. In the practical large-network implementation, the selector is the positive spectrum of the cross-entropy Hessian, used as a Fisher/GGN-motivated proxy. The longitudinal spine and the transverse subspace therefore play distinct, non-redundant roles: the spine provides low-loss centers around which posterior mass is placed, and the transverse sampling determines how it is dispersed. How much predictive diversity the longitudinal component adds depends on how much functional connectivity, rather than near-exact invariance, the traversed corridor actually contains.

Figure 1 in the main text illustrates the discrete support used by practical TRL. Figure 3 below focuses on the complementary frame-update view: updated tangents, transported transverse frames, and local perturbations around discrete spine anchors. This geometric mismatch with a single Euclidean ellipsoid is the starting point for TRL. The scalable method is a discrete transported low-rank approximation, not a fully smooth tubular pushforward.

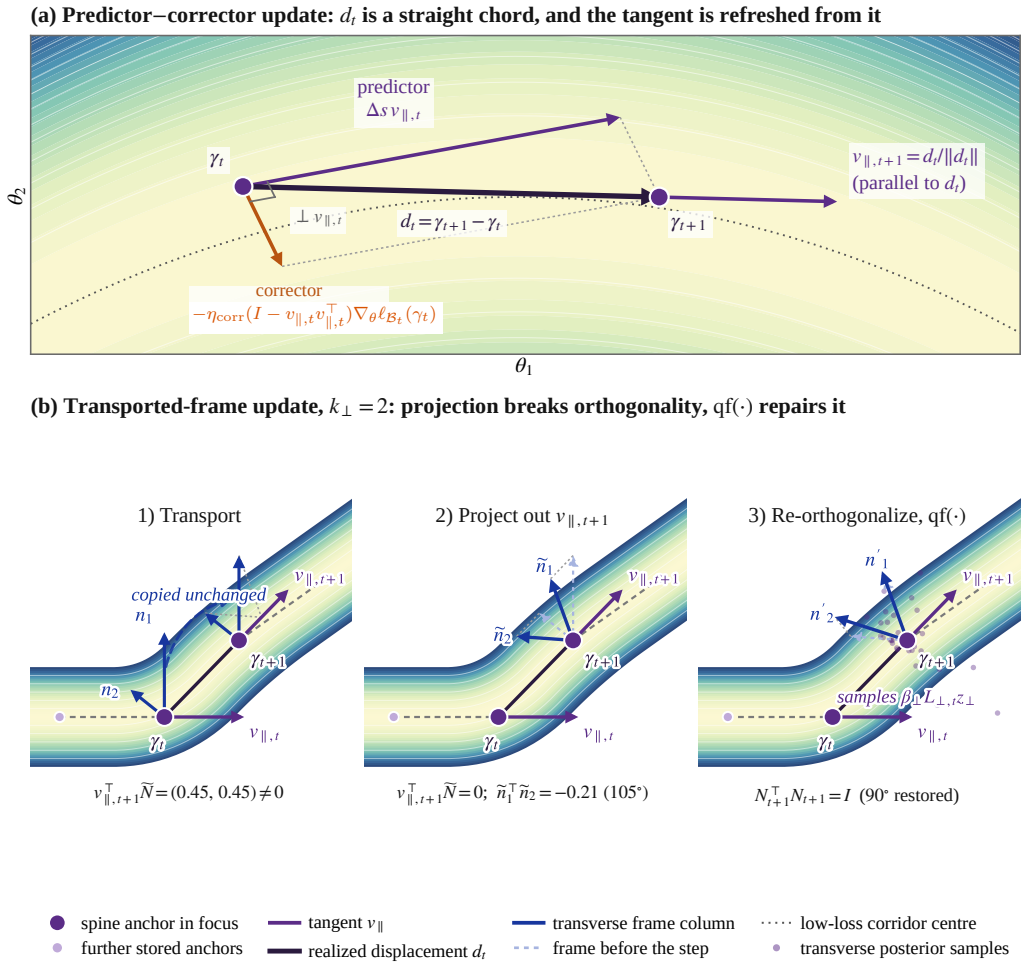


Figure 3: **Update mechanics of the discrete spine. (a)** One predictor–corrector step, Eq. (12). Both terms are evaluated at γ_t, so γ_{t+1} is computed rather than chosen: the predictor advances along the current tangent, which is only an approximate corridor direction, and the corrector projects out the component of $\nabla_\theta \ell_{\mathcal{B}_t}(\gamma_t)$ along that tangent and pulls back toward the low-loss region. The realized displacement $d_t = \gamma_{t+1} - \gamma_t$ therefore departs from $v_{\parallel,t}$, and Eq. (13) refreshes the tangent from d_t. The corrector-to-predictor ratio, the resulting tangent rotation, and the offset of γ_t from the corridor centre are all exaggerated for legibility; Appendix H reports the offsets actually observed along the spine. **(b)** The transported-frame update, Eq. (14), drawn with $k_\perp = 2$ because re-orthogonalisation is vacuous for a single column. Projecting the transported frame onto the complement of $v_{\parallel,t+1}$ removes the tangent component but leaves the columns mutually oblique ($\tilde{n}_1^\top \tilde{n}_2 = -0.21$, 105°); $\mathrm{qf}(\cdot)$ restores $N_{t+1}^\top N_{t+1} = I$ while preserving the direction of the first column. The frame is transported, not recomputed from a fresh local curvature decomposition at γ_{t+1}. Shading is the low-loss corridor, drawn from the same quadratic loss field in both panels. Figure 1 shows the resulting global support; this figure shows only the local update rule.

B.1 Fisher-motivated tangent/normal decomposition

The tangent/transverse interpretation follows the Fisher/GGN sensitivity picture in Appendix A: low-curvature directions approximate functionally weak or invariant motion, while leading positive-curvature directions define a data-sensitive transverse subspace. Practical TRL retains one constructed longitudinal direction and a rank-$k_\perp$ transverse basis rather than a full spectral decomposition. Figure 3 shows how that basis is transported and re-orthogonalized along the discrete spine.

B.2 PRIOR- VS DATA-DOMINATED VARIANCE

Let F_0 denote the likelihood-side Fisher/GGN curvature at the MAP, with eigenpairs $(u_i, \lambda_i(F_0))$, and let λI denote an isotropic prior precision. Under the idealized Gaussian precision $H \approx F_0 + \lambda I$, the variance along u_i is approximately $1/(\lambda_i(F_0) + \lambda)$. Thus:

- For tangential directions with $\lambda_i(F_0) \approx 0$, the variance is dominated by the prior, $\mathrm{Var}_{\parallel} \approx 1/\lambda$.

- For transverse directions with $\lambda_i(F_0) \gg \lambda$, the variance is dominated by the data, $\mathrm{Var}_{\perp} \approx 1/\lambda_i(F_0)$.

The tangential/transverse split above is most literal when transverse curvature dominates prior precision. In the low-rank subspace retained by the practical method, that condition holds only partially. The retained directions identify data-sensitive structure, but the HVP eigenvalues are computed from mini-batch-mean cross-entropy, whereas the inherited prior base precision comes from a Laplace marginal-likelihood fit that accumulates likelihood curvature over a 5,000-example subset. These terms are therefore not a matched dataset-sum Laplace precision, and the transverse Gaussian can be only weakly anisotropic and substantially prior-dominated. This separates support selection from scale: the Fisher/GGN picture motivates where transverse uncertainty should live, while the implementation selects that support with the positive spectrum of the cross-entropy Hessian and sets its magnitude largely through projected prior precision and validation tempering. The large-network sampler uses a two-block, block-isotropic approximation: classifier-head parameters use the base precision inherited from the last-layer Laplace fit, while non-head parameters use a boosted backbone precision. Appendix G.5 shows that this two-block structure is important for calibrated samples. Accordingly, $\mathrm{Var}_{\perp} \approx 1/\lambda_i(F_0)$ is a subspace-selection idealization rather than a literal description of the transverse covariance used by the large-network sampler.

A single Euclidean ellipsoid can become wide in poorly identified directions, but those directions need not remain aligned with a curved functional valley and may leave the low-loss corridor. Weight-space sampling can then place mass in directions that look flat under a local quadratic approximation but perform poorly under the nonlinear network. TRL instead allocates uncertainty to a low-rank tube: a Fisher/GGN-motivated positive-curvature proxy selects the transverse support, while the longitudinal component is anchored to a diagnosed low-loss spine. Appendix K makes this distinction explicit. Relative to Euclidean and linearized Laplace baselines, TRL changes where posterior mass is allocated—low-rank Gaussian uncertainty across selected data-sensitive directions plus anchored discrete coverage along a verified low-loss corridor—rather than changing the within-direction Laplace variance law.

B.3 CONTINUOUS GEOMETRIC IDEALIZATION

Figure 4 illustrates the continuous idealization: a latent coordinate controls displacement along a low-loss valley, while transverse coordinates control perturbations across it. The implementation discretizes this picture: samples are drawn around stored spine anchors and transverse perturbations are expressed in a transported low-rank basis. The idealization is explanatory rather than literal; it identifies the geometric objects of interest—a low-loss centerline, a transverse frame, and a local transverse scale—while the scalable algorithm specifies how they are approximated from one trained checkpoint.

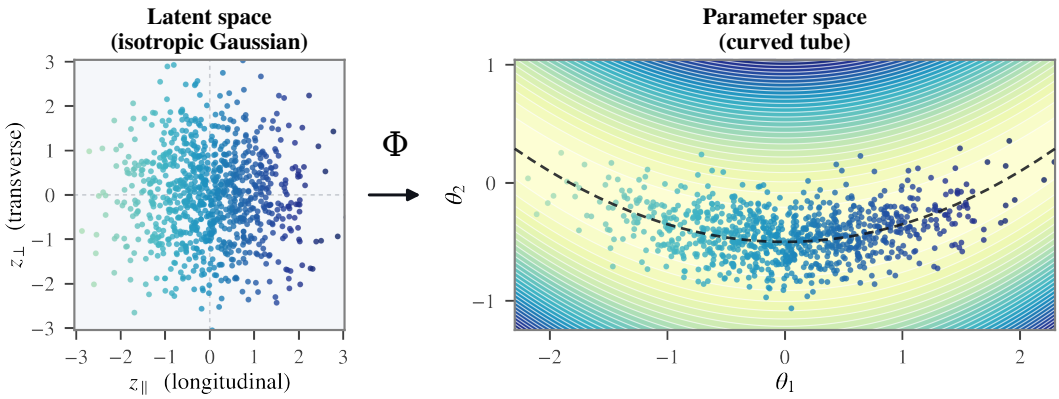


Figure 4: **Continuous tubular idealization.** An isotropic Gaussian in latent longitudinal/transverse coordinates is mapped by Φ to a curved tube in a schematic 2D parameter-space projection; θ_1 and θ_2 are projected coordinates, not individual weights. The dashed curve is the centerline and the spread represents local fibers. Practical TRL uses a finite predictor–corrector spine and transported low-rank transverse factors rather than a global smooth pushforward or repeated curvature decomposition.

Let $\theta_0 = \theta_{\mathrm{MAP}}$ and let $G(\theta)$ denote a positive semi-definite curvature metric, such as a Fisher or GGN approximation plus prior precision. The idealized TRL construction assumes a smooth curve

$$\gamma : \mathbb{R} \to \mathbb{R}^K, \qquad t \mapsto \gamma(t), \tag{16}$$

The curve in Eq. (16) satisfies $\gamma(0) = \theta_0$ and has approximately constant loss over the range of interest. Its tangent $\dot{\gamma}(t)$ represents a low-Fisher-curvature direction, while the transverse space captures directions in which the likelihood is data-sensitive.

At each point on the ideal curve, suppose we have a transverse frame $N(t) \in \mathbb{R}^{K\times k}$ and a transverse covariance square-root factor $L_\perp(t) \in \mathbb{R}^{k\times k}$. The idealized tubular map is

$$\Phi(z_\parallel, z_\perp) = \gamma(\alpha_\parallel z_\parallel) + N(\alpha_\parallel z_\parallel)L_\perp(\alpha_\parallel z_\parallel)z_\perp, \tag{17}$$

where $z = (z_\parallel, z_\perp) \sim \mathcal{N}(0, I_{1+k})$ and $\alpha_\parallel$ sets the longitudinal scale; the local transverse covariance in this idealization is $L_\perp L_\perp^\top$. This construction motivates TRL: posterior mass should be allowed to occupy flat, low-loss directions while remaining more cautious across data-sensitive transverse directions. Whether longitudinal averaging yields additional predictive diversity depends on whether the traversed corridor contains functional connectivity beyond near-exact invariances.

The map in Eq. (17) is a valid tubular parametrization only locally: its transverse radius must remain below the normal injectivity radius, or reach, of the curve γ. Beyond this radius, normal fibers can intersect and the coordinates $(z_\parallel, z_\perp)$ need not be unique. We use this condition only as part of the continuous geometric idealization. The practical sampler operates with a finite discrete spine and validation-selected transverse scales, but we do not explicitly certify the reach of the learned trajectory.

Equation (17) is a geometric model rather than the literal high-dimensional algorithm. A fully faithful implementation would require a smooth curve, continuous transport of frames, exact low-curvature tangents, and fresh local curvature factors along the entire path. That is computationally prohibitive in modern deep networks and is not the approximation used in our experiments. The implementation instead asks a more practical question: whether an amortized discrete approximation retains the parts of the geometry that matter empirically.

C Implementation and Reproducibility Details

This appendix specifies the experimental protocol and the practical implementation used in the consolidated ResNet experiments: concrete data splits, hyperparameter choices, baseline protocols, and BatchNorm recalibration procedure.

C.1 Dataset splits and validation protocol

CIFAR-100 experiments use ResNet-18-CIFAR with the standard CIFAR stem. We split the original training set into 45,000 training images and 5,000 held-out validation images; the standard 10,000-image test set is used only for final reporting. OOD AUROC uses SVHN as the out-of-distribution set for CIFAR-100 evaluations unless otherwise noted. Hyperparameters are selected exclusively on the clean validation split by minimizing validation NLL. In the consolidated CIFAR-100 experiments, all five seeds selected $\beta_\perp = 4.0$ from the grid $\{2, 3, 4, 6, 10, 20\}$. The same clean-validation-selected value is then reused for CIFAR-100-C; no corrupted validation data, corrupted test images, or test labels are used for tuning. Consolidated clean CIFAR-100 results aggregate five independently trained MAP checkpoints. CIFAR-100-C likewise aggregates three independently trained, seed-matched MAP checkpoints, whereas most fixed-MAP ablations use three posterior-sampling replicates on one checkpoint; the low-loss and functional-disagreement spine diagnostics use five independently trained checkpoint-specific runs, and the small-data fine-tuning study reports ten seeds. This protocol separates model selection from both clean test evaluation and corruption robustness evaluation. "Single-checkpoint post-hoc" means that TRL starts from one deterministically trained checkpoint, does not continue the original task training, and does not train additional independent backbones; it does not mean data-free post-processing. The construction uses training inputs for stochastic HVPs and, for BatchNorm networks, FixBN, while the clean held-out validation split is used for tube-scale and prior-boost selection.

C.2 Practical TRL algorithm

The practical sampler implements the discrete mixture of Appendix A.4. From a trained MAP checkpoint it estimates a leading positive-curvature rank-$k_\perp$ basis with stochastic cross-entropy HVPs, constructs a predictor–corrector spine in the complementary low-curvature region, transports and re-orthogonalizes the transverse basis, and samples a discrete anchor followed by a low-rank Gaussian perturbation. We treat this stochastic-HVP rank-$k_\perp$ basis as a scalable subsampled curvature proxy; it is not an exact full-data eigenspace: mini-batch quadratics can systematically distort extreme curvature directions relative to the full-data quadratic (Tatzel et al., 2025). A seed-0 exact-GGN selector replacement changed downstream Acc/NLL/ECE/Brier by only $-0.0014/-0.0025/+0.0008/-0.0009$, supporting predictive robustness without establishing eigenspace equivalence.

The main ResNet-18-CIFAR-100 setting uses $k_\perp = 30$, $T = 40$ anchors, $\Delta s = 0.01$, $S = 25$, $\beta_\perp = 4.0$, and 25 FixBN batches. Appendix G varies rank, anchor count, tube scale, and FixBN budget.

C.3 FixBN protocol

For architectures with BatchNorm, sampled weights can shift internal activation statistics. The reported CIFAR TRL tables use 25 FixBN mini-batches per posterior sample in *rolling* mode, in which running statistics are carried from one posterior draw into the next. The main ResNet-18 and WideResNet-16-4 SWAG-Diag rows use 20 posterior samples and 20 rolling-mode FixBN batches per sample. The VGG-11-BN SWAG-Diag row uses 25 posterior samples and 20 rolling-mode FixBN batches per sample. All reported SWAG-Diag runs are initialized from their seed-matched MAP checkpoints. In rolling mode, BatchNorm running-statistics buffers are not reset between sampled weight states. The validation sweep and subsequent final evaluation are therefore stateful: the running-statistics state reached during the selected sweep can carry into final prediction, creating the sweep-to-final dependency referenced in the Reproducibility Statement. MC Dropout recalibrates BatchNorm once before each predictive evaluation call and then performs its stochastic dropout passes without per-draw recalibration. In contrast, TRL and SWAG-Diag perform FixBN for each sampled weight state. The batch-count ablations in Appendix G use rolling mode and are labeled accordingly.

The released code defaults to independent *reset* recalibration. For every posterior draw, it calls `reset_running_stats()`, uses cumulative averaging with `momentum=None` during the recalibration passes, and restores the original momentum afterward. The ImageNet/ResNet-50 scale-check uses this reset protocol independently for every posterior sample; the reported CIFAR tables and their batch-count ablations use the rolling protocol described above. Architectures without running normalization statistics, including LayerNorm-based Vision Transformers, do not require FixBN; this

is an architectural consequence, not an unreported ViT experiment. Inputs are non-corrupted training inputs drawn from the training pipeline used for BN recalibration, including its standard preprocessing and augmentation. FixBN does not use labels, update weights, or access test/CIFAR-100-C images. It is an inference-time consistency correction after a weight sample has been generated; it does not define posterior support.

A paired seed-0 rolling-versus-reset audit found only small downstream changes for both TRL and SWAG-Diag and did not trigger a multi-seed rerun. A separate seed-0 end-to-end reconstruction matched the reported metric regime, whereas replay from the stored spine alone did not reproduce the stateful rolling-FixBN realization bitwise. These are implementation/provenance audits, not benchmark replicates; exact audit rows are retained in the release documentation.

C.4 Hyperparameter search spaces and selected values

For CIFAR-100 we use $k_\perp = 30$, $T = 40$, $\beta_\perp = 4.0$, $S = 25$, and 25 FixBN batches; $\beta_\perp$ is selected from $\{2, 3, 4, 6, 10, 20\}$ by held-out clean-validation NLL. The two-block isotropic prior assigns λ_{base} to the head and $\max(c\lambda_{\text{base}}, 5)$ to the backbone. We treat this as pragmatic regularization of the approximate posterior, not as a claim that the weight prior intrinsically splits into head and backbone components. The single boost coefficient c is likewise validation-selected; $c = 50$ is transferred unchanged to the architecture checks. Table 21 ablates c, showing the no-boost collapse and that c is not interchangeable with $\beta_\perp$.

C.5 Baseline configurations

This subsection distinguishes the baseline protocols used in the large-network CIFAR experiments from the full-network toy diagnostics. The method names are shared across settings, but the curvature approximation differs with scale.

The ResNet-18 CIFAR-100 benchmark defines the baseline protocol used throughout the main large-network experiments. MAP, ELA, LLA, SWAG-Diag, and TRL are attached to the same deterministic MAP checkpoints when applicable, whereas MC Dropout and Deep Ensembles require their own training-time stochastic or multi-seed training protocols. The WideResNet-16-4 and VGG-11-BN checks reuse these same baseline roles with architecture-specific checkpoints and validation-selected TRL tube scales; the VGG-11-BN SWAG-Diag row uses 25 posterior samples, compared with 20 in the ResNet-18 protocol, as documented above. The CIFAR-100→CIFAR-10 fine-tuning experiment is a separate diagnostic of spine behavior under task adaptation and is not part of the baseline suite.

Main ResNet-18 CIFAR-100 benchmark. The consolidated CIFAR-100 benchmark uses ResNet-18-CIFAR with the standard CIFAR stem. MAP is the deterministic model trained for 50 epochs with SGD and cosine scheduling. ELA and LLA use last-layer Laplace approximations with Kronecker-factored curvature and marginal-likelihood prior precision (Daxberger et al., 2021a; Immer et al., 2021a). The reported last-layer ELA/LLA fits use the same seeded 5,000-example subset used for the marginal-likelihood fit supplying λ_{base}; their curvature is accumulated on that subset without rescaling to the 45,000-example training set. These are the scalable ELA/LLA baselines reported in the main CIFAR tables: the posterior uncertainty is restricted to the classifier layer, and the difference between ELA and LLA is the predictive approximation used at test time. Deep Ensembles use $M = 5$ independently trained models and average their predictions (Lakshminarayanan et al., 2017). MC Dropout trains a dropout-modified ResNet-18 with dropout probability $p = 0.2$ and uses $S = 25$ stochastic dropout forward passes at inference (Gal & Ghahramani, 2016). The evaluated SWAG-Diag baseline is initialized from the MAP checkpoint and collects only the running per-parameter mean and second moment along a short additional fine-tuning trajectory; it does not store the low-rank snapshot-deviation matrix of full SWAG (Maddox et al., 2019). The published row uses 20 posterior samples and 20 rolling-mode FixBN batches per sample. A separate 20-versus-25 posterior-sample sensitivity check, holding FixBN fixed at 20 batches per draw, changed accuracy, NLL, and ECE by at most 0.0003, 0.000201, and 0.000326, respectively, across seeds 0–2; it does not replace the published 20/20 protocol.

Rank-matched all-weight Faller–Martin comparison. FM-LR-LLA ValScale is a checkpoint-matched integration of the predictive-covariance projector of Faller & Martin (2026). For each of

the five ResNet-18 checkpoints, we fit an all-weight Kronecker Laplace reference on 5,000 training examples with Laplace batch size 1, excluding BatchNorm affine parameters as in the reference configuration. We build the rank-30 projector from 256 examples and use 25 Monte Carlo samples from the induced linearized Gaussian logits. A scalar multiplier $\alpha_{\mathrm{FM}} \in \{0, 0.01, 0.03, 0.1, 0.3, 1\}$ rescales the projected covariance and is selected independently for each seed by clean-validation NLL; test and SVHN data are not used for selection. The $\alpha_{\mathrm{FM}} = 0$ sanity check reproduces the MAP probabilities exactly in every seed. Because covariance tempering and Monte Carlo linearized prediction strengthen the predictive baseline relative to the native unscaled/probit evaluation, we label the row ValScale to distinguish it from an unmodified reproduction. Appendix E.1 reports the per-seed results and projector diagnostics.

The random-subspace control of Appendix K replaces the HVP-selected transverse basis N_0 by a random rank-30 orthonormal basis (i.i.d. Gaussian draws orthonormalized by two modified Gram–Schmidt reorthogonalization passes in parameter space), keeping the block-isotropic prior, the validation grid for $\beta_\perp$, $S = 25$, and the FixBN protocol unchanged; we evaluate five independent bases on one MAP checkpoint and one basis on each of two further checkpoints.

Architecture-sensitivity checks. The WideResNet-16-4 and VGG-11-BN checks reuse the same CIFAR-100 data protocol, evaluation metrics, baseline roles, and clean-validation model selection rule, while retaining the stochastic budgets used by the architecture-specific runners. In particular, WideResNet-16-4 SWAG-Diag uses 20 posterior samples and 20 rolling-mode FixBN batches per sample, whereas the VGG-11-BN SWAG-Diag row uses 25 posterior samples and 20 rolling-mode FixBN batches per sample. They are intended to test whether the calibration behavior of TRL transfers across architectures, not to introduce new baseline definitions. Accordingly, MAP, SWAG-Diag, MC Dropout with $p = 0.2$, ELA, LLA, and TRL are implemented with the same roles as in the ResNet-18 benchmark, with architecture-specific checkpoint directories and classifier heads. The ELA/LLA baselines remain scalable last-layer Kronecker-factored Laplace baselines in these CIFAR-scale checks. Deep Ensembles are omitted from the supplementary WideResNet and VGG tables to focus on post-hoc and single-training methods under architecture changes.

Toy full-network diagnostics. The toy sine-regression and two-moons models are small enough that all-parameter, full-matrix GGN Laplace baselines are feasible. The released toy code uses `subset_of_weights="all"` and `hessian_structure="full"`; with the pinned `laplace-torch==0.2.2.2` default `CurvlinopsGGN` backend, "full" denotes a dense all-parameter curvature representation, not the exact Hessian of the training loss. In those tables, ELA and LLA use this dense GGN curvature, which isolates the geometric difference between a static local Gaussian and the TRL tube; the CIFAR-scale baselines use last-layer curvature. The toy results probe mechanism; the CIFAR-scale experiments carry the main evidence for scalable performance.

C.6 Runtime and memory measurement

Runtime in Table 14 is the sum of recorded timing keys, so scopes differ and nested keys can overlap. The MAP, ELA/LLA, and TRL rows come from the same five core executions; in each, the shared MAP stage took the training branch, so the 504.6 ± 7.2 s MAP time is already included in ELA/LLA and TRL. TRL additionally logs spine construction, the six-point validation sweep over $\beta_\perp$, and in-distribution/OOD posterior prediction. It excludes the preceding last-layer Laplace fit used to obtain λ_{base} and the separate validation sweep used to select $c = 50$. The ELA/LLA total exceeds MAP by 93.7 s, and that increment also contains prior optimization and four ELA/LLA prediction passes; hence the omitted λ_{base} fit alone is below 93.7 s, or less than 4.8% of the reported TRL total. FixBN is also logged both inside the enclosing prediction timers and as nested diagnostic sub-timers, inflating the summed TRL value. Thus 1985.8 s is an instrumented logged total with overlapping scopes, not a non-overlapping wall-clock measurement, and MAP must not be added to it again.

The Deep Ensemble row comes from separate recorded executions in which MAP was loaded and all five ensemble members were trained and evaluated; its 2444.1 s total therefore contains five member trainings, not a sixth MAP training. SWAG-Diag likewise loaded existing MAP checkpoints, explaining why its logged time can be below the standalone MAP-training row. MC Dropout mixes checkpoint hits and misses, which produces its large runtime variance. Peak VRAM is the maximum of the instrumented stages when available. These implementation- and hardware-dependent

totals characterize the observed pipelines and do not establish an absolute wall-clock ranking. The asymptotic comparison to ODE-based Riemannian Laplace samplers is given in Appendix L.

D TOY EXPERIMENTS WITH FULL-NETWORK LAPLACE

The toy experiments provide controlled settings in which all-parameter, full-matrix GGN Laplace baselines are tractable. They isolate the geometric effect of replacing a single static ellipsoid with a tubular approximation, without the layer restrictions and stochastic curvature approximations required by ResNet-scale experiments. These experiments are therefore diagnostic; the primary quantitative evidence is the consolidated CIFAR-100, CIFAR-100-C, and ablation results.

Purpose of the toy setting. The toys isolate posterior geometry with tractable dense all-parameter GGN curvature. Sine regression probes calibrated uncertainty under observation noise and substantial extrapolation; two-moons probes nonlinear classification. TRL initializes the longitudinal spine with the smallest-eigenvalue eigenvector of the full posterior-precision matrix at the MAP and updates subsequent tangents from normalized predictor–corrector displacements, while retaining a low-rank full-parameter transverse sampler. ELA and LLA use the same full-matrix GGN construction. Appendices D.1, D.2, and D.3 specify the regression setup, classification setup, and task-specific toy protocols.

D.1 NOISY SINE REGRESSION

For sine regression, 80 training inputs lie on an equally spaced grid over $[-3, 3]$, and the fixed 200-point evaluation grid spans $[-6, 6]$. Training and evaluation targets use independent noise realizations,

$$y_{\mathrm{tr}} = \sin(x_{\mathrm{tr}}) + \epsilon_{\mathrm{tr}}, \qquad y_{\mathrm{te}}^{\mathrm{obs}} = \sin(x_{\mathrm{te}}) + \epsilon_{\mathrm{te}}, \qquad \epsilon_{\mathrm{tr}}, \epsilon_{\mathrm{te}} \overset{\mathrm{ind}}{\sim} \mathcal{N}(0, 0.15^2). \tag{18}$$

Under the data model in Eq. (18), RMSE uses the latent $\sin(x_{\mathrm{te}})$; predictive NLL, standardized residuals, and coverage use $y_{\mathrm{te}}^{\mathrm{obs}}$. All probabilistic methods use the known observation-noise standard deviation $\sigma_{\mathrm{obs}} = 0.15$; Laplace prior precision is optimized independently for each training seed by marginal likelihood. For posterior mean μ_i and functional variance s_i^2, with $N = 200$ evaluation points, evaluation uses total predictive variance

$$\mathrm{NLL} = \frac{1}{N} \sum_i \frac{1}{2} \left[\log(2\pi v_i) + \frac{(y_i^{\mathrm{obs}} - \mu_i)^2}{v_i} \right], \qquad v_i = s_i^2 + 0.15^2. \tag{19}$$

With the predictive variance in Eq. (19), MAP retains the 0.15 observation-noise floor with $s_i^2 = 0$; functional variance excludes observation noise. We report $z_i = (y_i^{\mathrm{obs}} - \mu_i)/\sqrt{v_i}$, whose calibrated variance is 1; Cov. $1\sigma/2\sigma$ denote empirical one/two-standard-deviation coverage.

The longitudinal extent is selected without per-seed adaptation. We construct a 30-anchor candidate spine with $\Delta s = 0.04$, compare global prefixes of 10, 15, and 20 anchors by mean NLL over independent development seeds 50–69, and select one value for all subsequent runs. The 15-anchor prefix is selected and then frozen before evaluation on reporting seeds 100–199. The reporting set was first evaluated on 30 seeds and then extended to 100 without changing any hyperparameter. The transverse settings remain fixed at $\beta_\perp = 0.005$ and $k_\perp = 30$.

Table 4 reports the resulting 100-seed comparison. TRL improves mean NLL from 0.391 for LLA to 0.243. The paired difference $\mathrm{NLL}_{\mathrm{TRL}} - \mathrm{NLL}_{\mathrm{LLA}}$ is -0.149 with a 95% confidence interval of $[-0.204, -0.094]$, and TRL has lower NLL in 77 of 100 paired seeds. Its between-seed NLL standard deviation is larger than LLA's (0.260 versus 0.079), so we report the paired comparison rather than interpreting overlapping marginal error bars. RMSE remains comparable: the paired mean difference is -0.010 with interval $[-0.024, 0.003]$. Residual diagnostics show opposite scale errors: LLA is conservative ($\mathrm{Var}(z) = 0.574$, Cov. $2\sigma = 0.977$), while TRL remains mildly under-dispersed ($\mathrm{Var}(z) = 1.590$, Cov. $2\sigma = 0.884$).

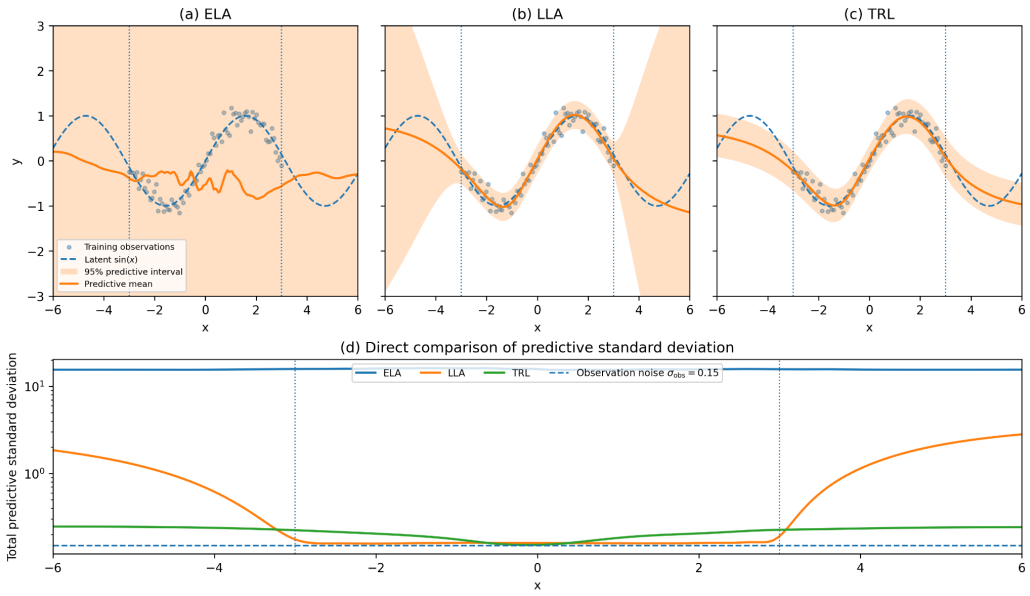


Figure 5: **Noisy sine regression under the frozen reporting protocol.** Top: predictive means and 95% predictive intervals for ELA, LLA, and TRL on a representative reporting seed selected deterministically as the seed whose TRL NLL is closest to the median over the 100 reporting seeds. Bottom: direct comparison of total predictive standard deviation on a logarithmic scale. The horizontal dashed line is the known observation-noise floor $\sigma_{\text{obs}} = 0.15$; vertical dotted lines delimit the training interval $[-3, 3]$. TRL uses the fixed global 15-anchor prefix.

Table 4: Noisy sine regression under the frozen final protocol. All entries are mean $\pm$ sample standard deviation over 100 reporting seeds. The known Gaussian observation-noise standard deviation is 0.15. Functional variance is estimated from posterior samples and excludes observation noise; deterministic MAP therefore has zero functional variance. TRL uses the globally selected fixed 15-anchor prefix.

Method	RMSE ↓	NLL ↓	$\mathrm{Var}(z)$	Cov. 1σ	Cov. 2σ	Func. var.
MAP	0.266 ± 0.047	1.138 ± 0.660	4.097 ± 1.289	0.480 ± 0.050	0.748 ± 0.052	0.00000 ± 0.00000
ELA	1.269 ± 0.351	3.673 ± 0.045	0.006 ± 0.004	1.000 ± 0.000	1.000 ± 0.000	246.6581 ± 22.0101
LLA	0.281 ± 0.070	0.391 ± 0.079	0.574 ± 0.101	0.824 ± 0.036	0.977 ± 0.011	1.70254 ± 0.64562
TRL	0.271 ± 0.038	0.243 ± 0.260	1.590 ± 0.702	0.598 ± 0.087	0.884 ± 0.075	0.03727 ± 0.02490

Figure 5 separates mean behavior from predictive scale. ELA is severely over-dispersed across the entire domain. LLA tracks the latent function near the training interval but its predictive standard deviation grows sharply under extrapolation. TRL keeps a moderate, input-dependent predictive scale and preserves a mean close to the trained predictor. The displayed seed is representative by a prespecified median-NLL rule; aggregate claims rely on Table 4 and the paired 100-seed comparison, not on visual selection.

D.2 Toy classification

For the two-moons experiment, we use a moderately noisy classification setting to avoid a saturated regime in which all methods achieve nearly perfect predictions. Specifically, we use noise 0.30, 500 training samples, and 1000 independently generated test samples. The generator noise perturbs the input coordinates and is not an observation-variance term added during evaluation. The classifier is a two-hidden-layer MLP with width 16 and a Bernoulli likelihood. Let $a_\theta(x)$ denote the scalar pre-sigmoid logit and $\sigma(a) = (1 + e^{-a})^{-1}$ the logistic sigmoid. The probabilities used for evaluation

are

$$p_i^{\mathrm{MAP}} = \sigma(a_{\theta_{\mathrm{MAP}}}(x_i))\,,$$
$$p_i^{\mathrm{TRL}} = \frac{1}{S}\sum_{r=1}^{S}\sigma(a_{\theta^{(r)}}(x_i))\,. \tag{20}$$

Both methods use the Bernoulli predictive NLL over the $N = 1000$ test points

$$\mathrm{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right], \tag{21}$$

We evaluate Eq. (21) with the corresponding MAP or posterior-averaged probability from Eq. (20), so posterior probabilities are averaged before taking the logarithm. This setting keeps the task nonlinear and nontrivial while still allowing dense all-parameter GGN computations. For this binary experiment, the Brier score uses the scalar Bernoulli convention $N^{-1}\sum_i (p_i - y_i)^2$, and functional variance is the posterior variance of the scalar probability $p(y = 1 \mid x)$. These are one-component binary quantities rather than the two-component vector convention used for multiclass classification; representing probabilities as $(1-p, p)$ would multiply both quantities by a factor of two.

Table 5 shows that TRL substantially improves NLL and Brier score over LLA, while deterministic MAP already matches TRL in accuracy, NLL, and Brier. Identical three-decimal MAP/TRL entries are a score tie, not identical predictive distributions: TRL has higher entropy and nonzero functional variance. Thus TRL adds a non-degenerate sampled posterior and functional variation, not a further score improvement over MAP in this regime. Entropy and functional variance are descriptive diagnostics, not monotone performance metrics.

Table 5: Two-moons classification with noise 0.30, 500 training points, and 1000 independent test points. Values are mean ± standard deviation over 10 seeds. Entropy and functional variance are diagnostics; deterministic MAP has zero functional variance.

Method	Acc. ↑	NLL ↓	Brier ↓	Entropy	Func. var.
MAP	0.910 ± 0.006	0.230 ± 0.014	0.067 ± 0.004	0.204 ± 0.007	0.0000 ± 0.0000
ELA	0.468 ± 0.086	0.704 ± 0.015	0.256 ± 0.007	0.690 ± 0.003	0.2002 ± 0.0038
LLA	0.904 ± 0.010	0.440 ± 0.029	0.133 ± 0.012	0.611 ± 0.018	0.1728 ± 0.0122
TRL	0.910 ± 0.006	0.230 ± 0.014	0.067 ± 0.004	0.216 ± 0.006	0.0012 ± 0.0009

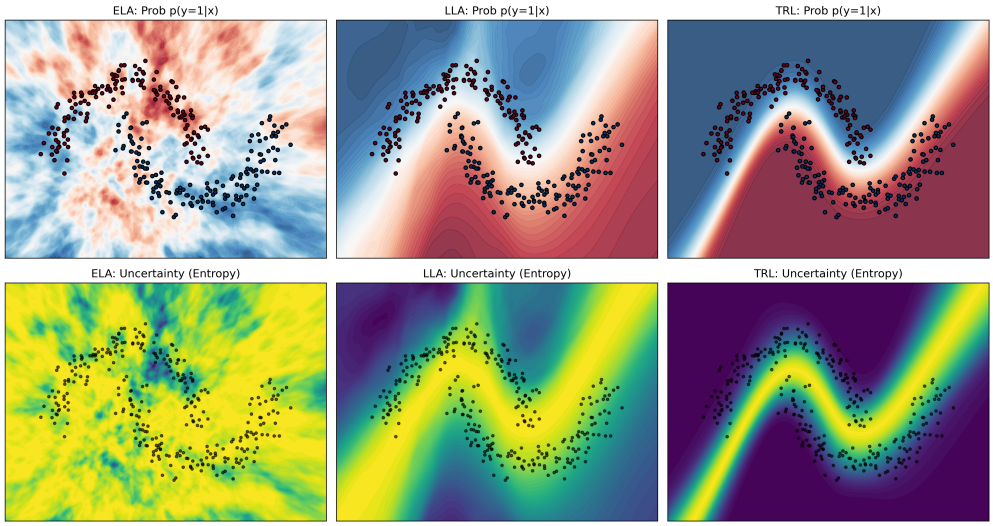


Figure 6: Two-moons classification. Predictive probability and entropy maps for ELA, LLA, and TRL.

Figure 6 illustrates the mechanism behind Table 5; deterministic MAP is reported only numerically. ELA remains high-entropy and weakly discriminative, while LLA captures the structure but is diffuse. TRL sharpens the boundary and improves NLL and Brier over LLA. Its scores remain nearly tied with MAP, so the additional output is the sampled posterior and nonzero functional variation. Relative to ELA, TRL also recovers a usable decision boundary.

The likelihoods make posterior uncertainty enter the two scores differently, but the observed contrast is also protocol-dependent. In Gaussian regression, functional variance appears explicitly in the predictive variance and can improve NLL even when the predictive mean is nearly unchanged; this effect is useful here because the sine test grid extends well beyond the training interval. In Bernoulli classification, posterior variation affects NLL only through the posterior-averaged class probability. In the present two-moons protocol, train and test data come from the same generator and MAP already yields strong predictive probabilities, leaving little room for posterior averaging to improve the reported scores.

D.3 Toy hyperparameters

The toy protocols are likelihood-specific but method-matched within each task. In sine regression, the known Gaussian observation-noise standard deviation $\sigma_{\text{obs}} = 0.15$ is supplied to the Laplace likelihood for ELA, LLA, and the Laplace object used by TRL; MAP is evaluated with the same observation-noise floor. Prior precision is optimized by marginal likelihood for each training seed. The candidate TRL trajectory has 30 anchors, $\Delta s = 0.04$, correction strength 0.10, $\beta_\perp = 0.005$, and $k_\perp = 30$. One global prefix length is selected from $\{10, 15, 20\}$ on development seeds 50–69 by mean NLL; the selected 15-anchor prefix is then fixed for all 100 reporting seeds 100–199.

Two-moons uses a Bernoulli likelihood, which has no scalar observation-noise parameter analogous to Gaussian σ_{obs}. We retain the prespecified classification configuration $T = 50$, $\Delta s = 0.08$, $\beta_\perp = 0.05$, and $k_\perp = 30$ without a spine-length search. This asymmetry is deliberate: the sine diagnostic is materially sensitive to longitudinal extent, whereas the two-moons full-spine and single-checkpoint variants are nearly tied in NLL, accuracy, and Brier score (Table 6), so additional classification tuning is not supported by the observed mechanism. ELA and LLA use all-parameter, full-matrix GGN Laplace approximations in both toys. Although dense all-parameter curvature is tractable here, TRL still uses a rank-30 transverse sampler: 30 of 2,701 parameters in the sine MLP and 30 of 337 in the width-16 two-moons MLP. Thus the toy sampler remains low-rank even though the underlying Laplace curvature can be materialized densely.

D.4 Isolating the longitudinal spine

To isolate longitudinal support, we compare the same local transverse sampler at the MAP against anchor mixing along the spine. For sine, "full-spine" denotes the globally selected fixed 15-anchor prefix; it is not selected separately for each reporting seed. For two-moons, the full prespecified 50-anchor spine is retained.

Table 6 shows a sharp task dependence. On sine regression, the fixed spine lowers NLL from 1.129 ± 0.651 to 0.243 ± 0.260 and wins all 100 paired seeds. The paired mean NLL difference is -0.887 with 95% interval $[-1.018, -0.756]$. RMSE remains statistically comparable: the full-minus-single difference is 0.005 with interval $[-0.004, 0.014]$. Functional variance rises from approximately 2.9×10^{-5} to 3.73×10^{-2}, showing that different anchors produce materially different functions of comparable point-prediction quality. On two-moons, full-spine and single-checkpoint variants remain nearly tied in accuracy, NLL, and Brier, so the transverse component dominates that classification diagnostic.

Table 6: Toy spine-isolation diagnostic. The sine rows use 100 frozen-configuration reporting seeds; TRL-FULL-SPINE denotes the fixed global 15-anchor prefix. The two-moons rows retain the prespecified configuration and use 10 seeds. Functional variance is a diagnostic, not a monotone score.

Task	Method	NLL ↓	RMSE ↓	Acc. ↑	Brier ↓	Func. var.
Sine	TRL-single-checkpoint	1.129 ± 0.651	0.266 ± 0.047	–	–	0.000029 ± 0.000011
Sine	TRL-full-spine	0.243 ± 0.260	0.271 ± 0.038	–	–	0.03727 ± 0.02490
Two-moons	TRL-single-checkpoint	0.230 ± 0.014	–	0.911 ± 0.005	0.067 ± 0.003	0.0009 ± 0.0005
Two-moons	TRL-full-spine	0.230 ± 0.014	–	0.910 ± 0.006	0.067 ± 0.004	0.0012 ± 0.0009

Overall, TRL improves over full-network ELA and LLA in both toys, but for different reasons. In sine regression, a globally fixed longitudinal prefix supplies decisive functional diversity and likelihood improvement while preserving comparable RMSE. In two-moons classification, MAP already achieves strong predictive scores, and the transverse sampler accounts for nearly all of TRL's improvement over the Laplace baselines. The result supports a task-dependent view of spine extent; longer longitudinal support is not universally beneficial.

E FULL CIFAR-100 RESULTS

Table 1 in the main text is the canonical five-checkpoint clean comparison. This appendix adds the matched all-weight low-rank, calibration-estimator, architecture-sensitivity, and runtime diagnostics needed to interpret that table.

E.1 RANK-MATCHED ALL-WEIGHT FALLER–MARTIN COMPARISON

The comparison in Table 7 targets the main potential confound of the last-layer Laplace rows: both FM-LR-LLA ValScale and TRL have access to the full network and use rank 30. The Faller–Martin projector is optimized to preserve the predictive covariance of an all-weight Kronecker Laplace reference on its construction set, whereas TRL selects parameter-space support with the positive-spectrum curvature proxy and evaluates nonlinear sampled networks. Table 7 matches checkpoint, architecture, split, and posterior rank, but not the prior family: FM uses a single scalar prior precision fitted for the Kronecker reference, while TRL uses the validation-selected two-block head/backbone precision described in Appendix C. It should therefore be read as a matched low-rank support-budget comparison, not as a one-factor ablation of the subspace criterion. Writing Ψ for the all-weight Kronecker-Laplace weight covariance and J_X for the stacked logit Jacobian on the 256-example projector set X, the predictive-covariance operator is $\Sigma_X = J_X \Psi J_X^\top$.

Table 7: FM-LR-LLA ValScale on CIFAR-100. The scalar covariance multiplier α_{FM} is selected independently per seed by clean-validation NLL from $\{0, 0.01, 0.03, 0.1, 0.3, 1\}$. Predictive metrics are test-set values. Trace fraction is the sum of the leading rank-30 eigenvalues divided by an eight-probe Hutchinson estimate of $\mathrm{tr}(\Sigma_X)$, where $\Sigma_X = J_X \Psi J_X^\top$ is the all-weight Kronecker predictive-covariance operator on the 256-example projector set. FM uses a scalar prior precision, whereas TRL uses a two-block head/backbone prior.

Seed	α_{FM}	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑	Trace frac.
0	0.03	0.7410	0.9913	0.0463	0.3582	0.8669	6.282%
1	0.10	0.7411	0.9857	0.0235	0.3603	0.8078	5.806%
2	0.03	0.7419	0.9883	0.0453	0.3620	0.8674	6.023%
3	0.03	0.7463	0.9845	0.0440	0.3555	0.8267	5.914%
4	0.03	0.7432	0.9744	0.0452	0.3535	0.8452	6.158%
Mean ± std.	–	0.7427 ± 0.0022	0.9848 ± 0.0064	0.0409 ± 0.0098	0.3579 ± 0.0034	0.8428 ± 0.0259	$6.037\% \pm 0.190\%$

Four of five seeds select $\alpha_{\mathrm{FM}} = 0.03$ and one selects 0.1; none selects the native-scale $\alpha_{\mathrm{FM}} = 1$, while $\alpha_{\mathrm{FM}} = 0$ exactly reproduces MAP probabilities. Thus the rank-30 all-weight covariance is predictively over-dispersed at native scale in this protocol; the $\alpha_{\mathrm{FM}} = 0$ sanity check rules out a mean-prediction or pipeline error. The retained projector captures $6.04\% \pm 0.19\%$ of the estimated predictive-covariance trace on its construction set. Relative to MAP, validation-scaled FM improves

NLL/ECE/Brier; TRL remains lower by $0.0336/0.0274/0.0045$ in the unrounded five-seed means and has 0.0243 higher mean predictive-entropy AUROC.

We report rank 30 as a matched support budget, not as an optimum for either method. FM rank truncates a specified predictive-covariance reference, whereas TRL rank is a support-design choice and its own rank ablation is non-monotone. Because a larger FM rank with covariance scale reselected at every rank is a distinct experiment that we did not run, the captured-trace fraction should not be interpreted as predicting the outcome of such a sweep.

E.2 Calibration sanity checks

Table 8 compares the headline five-checkpoint results with the current audited five-checkpoint MAP+TS control. The canonical checkpoint-matched top-label ECE comparison favors TRL in all five training seeds, with mean difference -0.0147 (TRL minus MAP+TS); a directly checkable sign certificate is given below. The three-seed clean MAP+TS result is retained only as the corruption-matched companion to the three-checkpoint CIFAR-100-C transfer artifact in Table 10; the five-checkpoint result is used for the clean headline comparison. The validation-NLL ELA/LLA grid-search results are reported separately below.

Table 8: CIFAR-100 calibration sanity checks. Values are mean $\pm$ sample standard deviation over five independently trained MAP checkpoints. ECE is 15-bin equal-width top-label ECE.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓	prior / temp.
MAP	0.743 ± 0.003	1.040 ± 0.008	0.096 ± 0.002	0.370 ± 0.003	–
MAP+TS	0.743 ± 0.003	0.960 ± 0.006	0.028 ± 0.004	0.355 ± 0.003	T=1.378 ± 0.026
ELA-marglik	0.724 ± 0.002	1.356 ± 0.013	0.262 ± 0.008	0.468 ± 0.004	$\lambda \approx 4.9$
LLA-marglik	0.739 ± 0.002	1.492 ± 0.017	0.368 ± 0.008	0.521 ± 0.006	$\lambda \approx 4.9$
TRL	0.744 ± 0.003	0.951 ± 0.006	0.013 ± 0.002	0.353 ± 0.003	$\beta_\perp = 4,\ c = 50$

Aggregate-only grid-search diagnostic. Using the same held-out validation split and a 100-point logarithmic grid $\lambda \in [10^{-4}, 10^4]$, the three-seed aggregate means were ELA-grid: Acc. 0.742, NLL 1.043, ECE 0.0956, Brier 0.3715; and LLA-grid: Acc. 0.742, NLL 1.042, ECE 0.0958, Brier 0.3716. All three seeds selected the upper boundary $\lambda = 10^4$ for both variants. Because the released artifact does not retain the corresponding per-seed metric rows, these aggregate-only values are reported in text and omitted from the uncertainty-aware rows in Table 8.

Validation-NLL tuning of last-layer Laplace. Under marginal-likelihood tuning, ELA/LLA select a small prior precision ($\lambda \approx 4.9$) and are strongly overdispersed, giving worse NLL and ECE than the MAP predictor itself. Re-tuning the prior precision by validation-NLL grid search–the same selection criterion used for TRL's tube scale-removes this overdispersion, but the selected value is the upper boundary of the search grid ($\lambda = 10^4$) for every seed and for both predictive variants. In other words, validation-NLL tuning does not recover a stronger non-degenerate last-layer posterior; it drives the posterior toward the MAP predictor. This is why ELA-grid and LLA-grid are nearly identical despite using different predictive approximations (probit GLM versus Monte Carlo nonlinear prediction): once the posterior collapses onto the MAP point, there is negligible weight-space variance to propagate, so the link approximation has little effect. Because this sensitivity varies only the isotropic prior precision, it does not emulate refitting the likelihood curvature on all 45,000 training examples. This control therefore addresses prior-selection sensitivity within the reported subset-fit last-layer Laplace protocol while showing that, in this CIFAR-100 regime, the tested last-layer Laplace family does not provide a useful calibration improvement over MAP unless it degenerates toward the deterministic predictor.

Temperature scaling. MAP temperature scaling uses one positive scalar temperature per seed, selected by NLL on the same clean validation split and frozen before test evaluation. The current five-checkpoint rerun selects $\tau = 1.378 \pm 0.026$. Writing a for the logit vector, the map $a \mapsto a/\tau$ is order-preserving for $\tau > 0$ and therefore preserves the checkpoint-matched MAP argmax exactly. It reduces mean NLL from about 1.040 to 0.960, 15-bin equal-width top-label ECE from 0.096 to 0.028, and Brier score from 0.370 to 0.355. TRL remains lower at $0.951/0.013/0.353$; all five paired checkpoint-wise ECE differences are negative, with mean difference -0.0147 (TRL minus

MAP+TS), while the NLL and Brier mean gaps are 0.0092 and 0.0018. We report temperature scaling as a calibration-only control, not as a posterior method: it returns one deterministic rescaled predictive distribution, whereas TRL returns parameter-space posterior samples and functional-variation diagnostics. MAP+TS clean predictive-entropy AUROC against SVHN is 0.870 ± 0.022.

Seedwise sign certificate. The canonical TRL top-label ECEs for seeds 0–4 are 0.014815, 0.010418, 0.015262, 0.013676, and 0.013025. For any $n = 5$ observations with mean μ and sample standard deviation s, each observation obeys $|x_i - \mu| \leq \sqrt{n-1}\, s$. The preserved MAP+TS aggregate gives $\mu = 0.028096$ and $s = 0.003533$; even allowing for their six-decimal display rounding, every MAP+TS seed therefore satisfies $x_i \geq 0.0210285$. This exceeds the largest canonical TRL seed ECE, 0.015262, so all five checkpoint-matched TRL-minus-MAP+TS differences are negative. This certificate establishes the signs from the preserved aggregate and canonical TRL rows; it does not reconstruct the unpreserved per-seed MAP+TS magnitudes. The machine-readable certificate is included as `ece_sign_certificate.csv` in the MAP+TS provenance directory; its schema separates per-seed upper bounds from the aggregate mean difference.

Table 9: Sensitivity of clean calibration conclusions to the estimator (mean $\pm$ sample standard deviation over three checkpoint-specific evaluations, training seeds 0–2). TRL here is the fresh-refresh mechanism diagnostic because that artifact preserves per-example probabilities; the headline five-checkpoint row remains canonical. Equal-mass and equal-width columns are top-label ECE with 15 bins. Lower is better.

Method	Classwise ECE	Equal-mass-15	Equal-width-15	KS calibration distance
Deep Ensemble	0.001743 ± 0.000030	0.01705 ± 0.00221	0.01775 ± 0.00156	0.01266 ± 0.00161
MAP	0.002619 ± 0.000015	0.09641 ± 0.00269	0.09641 ± 0.00269	0.09641 ± 0.00269
MAP+TS	0.001921 ± 0.000004	0.02834 ± 0.00450	0.02946 ± 0.00402	0.01703 ± 0.00678
TRL fresh-refresh	0.001960 ± 0.000024	0.01283 ± 0.00384	0.01268 ± 0.00329	0.00545 ± 0.00178

Table 9 supports a narrow but stable conclusion. Under both top-label binning rules and the bin-free KS distance, the fresh-refresh TRL diagnostic is lower than MAP+TS; the three paired equal-width differences (TRL minus MAP+TS) are all negative, with mean -0.0168 ± 0.0070. The conclusion is not universal calibration dominance: classwise ECE is marginally lower for MAP+TS, and Deep Ensembles are numerically best in the classwise column. The fresh-refresh seed-0 top-label ECE differs from the canonical seed-0 value by only about 0.00045, so the variant label is important for provenance but not material to the clean top-label conclusion. We therefore use "top-label ECE" in headline claims and retain NLL and Brier as the primary proper scoring rules.

Frozen temperature scaling under corruption. We additionally fit one scalar temperature per seed by minimizing NLL on the clean 5,000-example validation split, freeze it, and reuse it for the clean test set and all 19×5 CIFAR-100-C corruption/severity cells. No corruption data are used for calibration or model selection. MAP+TS and TRL use the same matched MAP checkpoints; the checkpoint identities were verified by SHA-256 and by exact equality of MAP and MAP+TS argmax accuracies in all 285 corruption cells. Both methods are selected by clean-validation NLL and transferred unchanged to CIFAR-100-C, although MAP+TS performs continuous optimization of one scalar whereas TRL selects its transverse scale from a discrete grid with Monte Carlo predictions. The anonymous supplementary result file contains 3 clean rows and 285 corruption rows with no missing or duplicate cells. Its clean MAP+TS row is the three-checkpoint companion to this frozen-corruption evaluation and should not be confused with the current five-checkpoint clean headline row in Tables 1 and 8.

Table 10: Frozen clean-validation temperature scaling for the three-checkpoint corruption transfer. CIFAR-100-C values macro-average the 95 corruption/severity cells within each of three independently trained, seed-matched checkpoints, then report mean ± sample standard deviation. The clean MAP+TS row is the corruption-matched three-checkpoint companion; the current five-checkpoint clean headline row is reported elsewhere. The clean TRL row reproduces the five-checkpoint headline result, while the CIFAR-100-C TRL row uses the three matched checkpoints.

Dataset	Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓
Clean (3-checkpoint companion)	MAP+TS	0.7415 ± 0.0014	0.9635 ± 0.0042	0.0296 ± 0.0042	0.3570 ± 0.0017
Clean	TRL	0.744 ± 0.003	0.951 ± 0.006	0.013 ± 0.002	0.353 ± 0.003
CIFAR-100-C	MAP+TS	0.4711 ± 0.0044	2.3895 ± 0.0411	0.1180 ± 0.0078	0.6907 ± 0.0070
CIFAR-100-C	TRL	0.473 ± 0.004	2.464 ± 0.047	0.118 ± 0.006	0.689 ± 0.005

Table 11: CIFAR-100-C calibration-only control by severity. MAP+TS is macro-averaged over 19 corruptions within each of three independently trained, seed-matched checkpoints, followed by mean ± sample standard deviation. TRL values are reproduced from Table 15.

Severity	Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓
1	MAP+TS	0.6210 ± 0.0048	1.5340 ± 0.0332	0.0570 ± 0.0068	0.5087 ± 0.0071
	TRL	0.623 ± 0.005	1.540 ± 0.030	0.047 ± 0.005	0.506 ± 0.005
2	MAP+TS	0.5330 ± 0.0053	1.9717 ± 0.0439	0.0836 ± 0.0084	0.6142 ± 0.0080
	TRL	0.535 ± 0.004	1.998 ± 0.041	0.079 ± 0.007	0.611 ± 0.006
3	MAP+TS	0.4774 ± 0.0045	2.3039 ± 0.0434	0.1094 ± 0.0074	0.6827 ± 0.0067
	TRL	0.480 ± 0.004	2.352 ± 0.050	0.108 ± 0.005	0.679 ± 0.005
4	MAP+TS	0.4104 ± 0.0036	2.7385 ± 0.0438	0.1437 ± 0.0081	0.7650 ± 0.0068
	TRL	0.412 ± 0.003	2.841 ± 0.059	0.148 ± 0.003	0.763 ± 0.003
5	MAP+TS	0.3136 ± 0.0042	3.3994 ± 0.0482	0.1964 ± 0.0107	0.8832 ± 0.0089
	TRL	0.315 ± 0.004	3.589 ± 0.066	0.209 ± 0.009	0.886 ± 0.009

The corruption NLL advantage is consistent across the three paired seeds. Defining $\Delta = \mathrm{NLL}_{\mathrm{TRL}} - \mathrm{NLL}_{\mathrm{MAP+TS}}$, the differences are 0.086415, 0.075855, and 0.061099, giving 0.074456 ± 0.012716. With only three independent seeds, we treat this as consistent descriptive evidence; three independent seeds are insufficient for a strong statistical claim. The result delimits the calibration claim: scalar recalibration is a strong corruption-NLL control, but it does not define a posterior over parameters or functions, generate posterior samples, or provide the support and functional-variation diagnostics available from TRL.

E.3 Architecture-sensitivity check: WideResNet-16-4

To test whether the calibration behavior is tied to a single architecture, we also run a supplementary CIFAR-100 check with WideResNet-16-4-CIFAR. This is not intended as a full architecture-scaling study. We reuse the main TRL structural configuration and select only the tube scale on the clean validation split using the same validation-NLL criterion and grid as in the ResNet-18 experiments. The selected value is $\beta_\perp = 3.0$ for all three WideResNet seeds, whereas the ResNet-18 experiments select $\beta_\perp = 4.0$. This difference is expected because the effective posterior scale depends on architecture-specific curvature and normalization behavior.

Table 12: Supplementary architecture-sensitivity check on CIFAR-100 with WideResNet-16-4-CIFAR. Values are mean ± standard deviation over three independent training seeds per method. This table compares single-training baselines; Deep Ensembles are omitted from this supplementary check to focus on the architecture sensitivity of post-hoc and single-training methods.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑
MAP	0.739 ± 0.002	1.026 ± 0.017	0.095 ± 0.002	0.373 ± 0.004	0.798 ± 0.021
SWAG-Diag	0.740 ± 0.001	1.032 ± 0.014	0.097 ± 0.001	0.373 ± 0.004	0.798 ± 0.018
MC Dropout	0.747 ± 0.002	0.910 ± 0.005	0.053 ± 0.002	0.350 ± 0.001	0.839 ± 0.020
ELA (last-layer)	0.728 ± 0.002	1.124 ± 0.010	0.150 ± 0.003	0.407 ± 0.003	0.852 ± 0.021
LLA (last-layer)	0.729 ± 0.002	1.172 ± 0.010	0.207 ± 0.006	0.428 ± 0.003	0.888 ± 0.017
TRL	0.740 ± 0.001	0.944 ± 0.007	0.028 ± 0.004	0.359 ± 0.002	0.817 ± 0.029

Table 12 shows a mixed WideResNet-16-4 result. TRL substantially improves over MAP, SWAG-Diag, ELA, and LLA in NLL, ECE, and Brier score, and has the best ECE among the evaluated methods; MC Dropout is stronger in accuracy, NLL, and Brier. The calibration benefit therefore transfers beyond ResNet-18, but the ranking among single-training methods changes with architecture and regularization. Appendix E.4 reports a second check on a non-residual network.

Seed-count note. The clean headline table aggregates five independently trained checkpoints; CIFAR-100-C and the architecture checks use three checkpoint-specific runs; fixed-MAP ablations use posterior-sampling replicates unless their captions state otherwise; the boost-prior ablation uses three independently trained checkpoints. Transport and fine-tuning diagnostics use the seed counts stated locally. These error bars therefore quantify different sources of variability and should be compared only within the corresponding study.

E.4 Architecture-sensitivity check: VGG-11-BN

As a second architecture check, we run CIFAR-100 with VGG-11-BN-CIFAR. VGG-11-BN has no residual skip connections. Because residual structure contributes some of the near-exact functional invariances that shape low-loss valleys, this non-residual topology provides a useful stress test of whether TRL's calibration behavior transfers beyond residual networks. We reuse the main TRL structural configuration ($k_\perp = 30$, $T = 40$, $S = 25$, 25 FixBN batches) and select only the tube scale on the clean validation split by validation NLL. All three seeds select $\beta_\perp = 4.0$, the same value selected by the ResNet-18 experiments, with a clear interior optimum on the validation-NLL curve; this matches the WideResNet selection ($\beta_\perp = 3.0$) to within one grid step and indicates that the transverse tempering scale is not strongly architecture-sensitive across the three architectures studied.

Table 13: Supplementary architecture-sensitivity check on CIFAR-100 with VGG-11-BN-CIFAR. Values are mean ± standard deviation over three independent training seeds per method. As in the WideResNet-16-4 check, this table compares single-training baselines and omits Deep Ensembles. Among single-checkpoint post-hoc methods (MAP, ELA, LLA, and TRL), TRL gives the best ECE, NLL, and Brier. SWAG-Diag starts from MAP but collects statistics along a short additional fine-tuning trajectory. MC Dropout, which acquires its uncertainty during training rather than post-hoc (Section 5), is the strongest single-training baseline on this architecture.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑
MAP	0.678 ± 0.002	1.377 ± 0.011	0.134 ± 0.002	0.462 ± 0.003	0.780 ± 0.010
SWAG-Diag	0.679 ± 0.002	1.377 ± 0.010	0.135 ± 0.001	0.461 ± 0.003	0.778 ± 0.007
MC Dropout	0.686 ± 0.001	1.135 ± 0.005	0.023 ± 0.001	0.419 ± 0.003	0.808 ± 0.029
ELA (last-layer)	0.656 ± 0.003	1.587 ± 0.008	0.220 ± 0.007	0.528 ± 0.001	0.815 ± 0.009
LLA (last-layer)	0.674 ± 0.002	1.670 ± 0.012	0.324 ± 0.007	0.568 ± 0.003	0.852 ± 0.009
TRL	0.676 ± 0.003	1.248 ± 0.014	0.028 ± 0.003	0.438 ± 0.005	0.803 ± 0.015

Table 13 shows that the VGG-11-BN result is also mixed. Among the single-checkpoint post-hoc methods (MAP, ELA, LLA, and TRL), TRL is the strongest: it reduces ECE from 0.134 ± 0.002

(MAP) to 0.028 ± 0.003, roughly a five-fold improvement, and similarly improves NLL and Brier over MAP, ELA, and LLA. It also improves these three metrics over MAP-initialized SWAG-Diag, whose three-seed result is NLL 1.377 ± 0.010, ECE 0.135 ± 0.001, and Brier 0.461 ± 0.003. The SWAG-Diag predictive-entropy AUROC is 0.778 ± 0.007, below TRL's 0.803 ± 0.015, while LLA retains the highest AUROC in the table. On this non-residual network, TRL again yields a markedly better-calibrated posterior predictive than the static local-Gaussian and diagonal checkpoint-initialized baselines.

The comparison with MC Dropout is the one point where this architecture departs from the ResNet-18 ordering. On VGG-11-BN, MC Dropout is better than TRL in accuracy, NLL, ECE, and Brier, with non-overlapping three-seed intervals. Because dropout is introduced during training, this comparison does not isolate post-hoc inference alone. Nevertheless, the result exposes a limitation of TRL on this architecture: the deterministic VGG checkpoint does not yield as effective a post-hoc uncertainty model as training-time dropout. Both methods use $S = 25$ and follow the method-specific normalization procedures documented in Appendix C. TRL still improves over the evaluated post-hoc baselines, but the ordering against training-time stochastic methods is architecture-dependent.

E.5 Runtime and memory breakdown

Table 14: Instrumented CIFAR-100 logged-time and peak-VRAM summary. Mean $\pm$ standard deviation over five seeds. The time column is the sum of recorded timing keys; scopes differ by method, and the TRL row contains nested FixBN sub-timers already included in its prediction timers. It is therefore not a normalized, non-overlapping wall-clock benchmark.

Method	Logged time (s)	Peak VRAM (GB) ↓
MAP	504.6 ± 7.2	0.70 ± 0.00
ELA / LLA (shared MAP train + fit + pred)	598.3 ± 7.3	9.92 ± 0.00
SWAG-Diag	192.3 ± 108.5	0.66 ± 0.29
MC Dropout	277.5 ± 292.1	0.51 ± 0.34
Deep Ensemble	2444.1 ± 50.4	0.79 ± 0.00
TRL	1985.8 ± 31.4	6.42 ± 0.00

Table 14 documents the observed pipeline instrumentation; it is not a normalized wall-clock benchmark. TRL uses one independently optimized backbone but adds substantial post-hoc construction, validation-scale selection, and sampled-network inference. Deep Ensembles use five independently optimized backbones. Because the recorded totals differ in scope and the TRL total contains overlapping nested timers, we make no absolute wall-clock ranking between them. The supported computational distinction is the number and role of the optimization runs, not a claim that the displayed TRL total is intrinsically lower. The runtime columns in Appendix G are separate study-local measurements and must also be compared only within their own tables.

F CIFAR-100-C and Human-Ambiguity Diagnostics

CIFAR-100-C evaluates corruption robustness without changing the trained model or tuning hyperparameters on corrupted data. This setting is useful for separating deterministic accuracy from probabilistic reliability: under stronger corruption, all methods lose accuracy, but posterior averaging and scalar recalibration can affect NLL, ECE, and Brier differently. The clean-validation-selected TRL configuration is used unchanged. Table 2 in the main text retains the posterior/ensemble comparison; the separately audited MAP+TS calibration control is reported in Tables 10–11.

Table 15: CIFAR-100-C by severity for Deep Ensembles, MC Dropout, and TRL. Mean ± sample standard deviation over three independently trained, seed-matched checkpoints; each TRL posterior is constructed from its corresponding MAP checkpoint.

Severity	Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓
1	Deep Ensemble	0.663 ± 0.001	1.358 ± 0.004	0.045 ± 0.001	0.4565 ± 0.0005
	MC Dropout	0.624 ± 0.003	1.554 ± 0.012	0.094 ± 0.004	0.511 ± 0.004
	TRL	0.623 ± 0.005	1.540 ± 0.030	0.047 ± 0.005	0.506 ± 0.005
2	Deep Ensemble	0.575 ± 0.001	1.811 ± 0.004	0.075 ± 0.001	0.565 ± 0.001
	MC Dropout	0.538 ± 0.005	2.035 ± 0.039	0.131 ± 0.007	0.621 ± 0.008
	TRL	0.535 ± 0.004	1.998 ± 0.041	0.079 ± 0.007	0.611 ± 0.006
3	Deep Ensemble	0.517 ± 0.002	2.183 ± 0.012	0.105 ± 0.001	0.639 ± 0.003
	MC Dropout	0.482 ± 0.006	2.423 ± 0.074	0.162 ± 0.010	0.696 ± 0.012
	TRL	0.480 ± 0.004	2.352 ± 0.050	0.108 ± 0.005	0.679 ± 0.005
4	Deep Ensemble	0.447 ± 0.002	2.677 ± 0.018	0.143 ± 0.003	0.729 ± 0.005
	MC Dropout	0.413 ± 0.006	2.974 ± 0.087	0.206 ± 0.010	0.790 ± 0.013
	TRL	0.412 ± 0.003	2.841 ± 0.059	0.148 ± 0.003	0.763 ± 0.003
5	Deep Ensemble	0.344 ± 0.002	3.429 ± 0.026	0.201 ± 0.004	0.859 ± 0.005
	MC Dropout	0.317 ± 0.008	3.785 ± 0.086	0.268 ± 0.010	0.920 ± 0.014
	TRL	0.315 ± 0.004	3.589 ± 0.066	0.209 ± 0.009	0.886 ± 0.009

Table 2 in the main text reports the aggregate CIFAR-100-C summary, while Table 15 captures the degradation trend across corruption strengths. Deep Ensembles dominate the posterior/ensemble table at every severity, while TRL improves the probabilistic metrics over MC Dropout and remains close to the ensemble ECE at mild and moderate corruptions. The calibration-only MAP+TS control has lower NLL than TRL at each severity; ECE and Brier are close, with ordering varying by metric and severity. At severity 5, the gap between within-basin methods and Deep Ensembles widens, reflecting the limitation of a single-basin approximation under strong distribution shift. TRL is therefore best interpreted as a posterior approximation that remains competitive with scalar recalibration while providing posterior samples and functional diagnostics, not as a universal replacement for either temperature scaling or cross-basin ensemble diversity.

F.1 CIFAR-10H HUMAN-LABEL ALIGNMENT

CIFAR-10H provides an empirical human label distribution $q_H(\cdot \mid x)$ for each of the 10,000 CIFAR-10 test images, with roughly 51 human judgments per image on average (Peterson et al., 2019). We use these distributions only for evaluation: no CIFAR-10H labels enter training, posterior construction, FixBN, sample-budget choice, or hyperparameter selection. The official row order was independently checked against the raw CIFAR-10H test indices and true labels; all 10,000 rows match the CIFAR-10 test ordering used by our saved predictions. We evaluate three independently trained ResNet-18 seeds from the stored CIFAR-10 runs. Each deterministic MAP checkpoint was trained from scratch for 50 epochs with SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}), cosine annealing, batch size 128, a 45,000/5,000 train/validation split, and standard random-crop plus horizontal-flip augmentation. Deep-Ensemble members use the same 50-epoch training recipe; MC Dropout uses the corresponding dropout-trained model.

For the CIFAR-10H evaluation, MAP is deterministic and the Deep Ensemble averages five members within each seed-group. MC Dropout uses 25 stochastic predictive passes after the one-time 20-batch BatchNorm recalibration. TRL replays each frozen posterior/spine with the validation-selected $\beta_\perp$ stored in that payload (`best_tube_scale`); the associated validation sweep selected $\beta_\perp = 2.0$ in each of seeds 0, 1, and 2. No HVP, spine reconstruction, or tube-scale re-selection is performed. We draw 25 fresh posterior samples with a fixed evaluation seed and apply the 25-batch *rolling* FixBN procedure to each draw (running statistics are not reset between draws). SWAG-Diag is collected from the corresponding MAP checkpoint with 10 fine-tuning epochs at learning rate 10^{-3}, unbiased arithmetic diagonal moments, sample scale 1.0, 20 posterior draws, and independent 20-batch reset FixBN. CIFAR-10 SWAG caches initialized from an ensemble member are not used. These method-specific FixBN semantics reflect artifact provenance, not CIFAR-10H tuning: the frozen TRL replay

retains its rolling protocol, whereas SWAG-Diag uses the reset protocol. A separate paired seed-0 CIFAR-100 audit (Appendix C.3) found only small rolling-versus-reset changes for both methods and did not trigger reruns; this is a sensitivity/provenance audit, not a matched CIFAR-10H recalibration experiment. Predictions are evaluated separately for each training seed and only the resulting metrics are aggregated across seeds; predictions are never averaged across the three training seeds.

For each image we compare the model predictive distribution p with q_H using $\mathrm{KL}(q_H \| p)$, soft-target Brier score, Jensen–Shannon (JS) divergence, signed entropy bias $H(p) - H(q_H)$, and Pearson correlation between predictive and human entropy across images. Table 16 reports the full 10,000-image test set and the 5,607-image subset with nonzero human-label entropy. The latter isolates images on which the human annotations exhibit perceptual ambiguity.

Table 16: CIFAR-10H alignment (mean $\pm$ sample SD over three independently trained seeds). Lower is better for KL, Brier, and JS; entropy bias is signed with zero ideal; higher entropy correlation is better. "Ambiguous" means $H(q_H) > 0$ (5,607/10,000 images). No CIFAR-10H labels are used for training or tuning.

Subset	Method	KL ↓	Brier ↓	JS ↓	H bias (0 ideal)	H corr. ↑
All	MAP	0.3819 ± 0.0180	0.0857 ± 0.0049	0.0422 ± 0.0017	-0.0687 $\pm$ 0.0026	0.2827 ± 0.0136
	SWAG-Diag	0.3894 ± 0.0190	0.0853 ± 0.0054	0.0420 ± 0.0019	-0.0722 $\pm$ 0.0023	0.2796 ± 0.0113
	MC Dropout	0.3026 ± 0.0051	0.0769 ± 0.0029	0.0406 ± 0.0011	-0.0260 $\pm$ 0.0028	0.3290 ± 0.0064
	TRL (ours)	0.2939 ± 0.0110	0.0776 ± 0.0044	0.0434 ± 0.0018	+0.0243 $\pm$ 0.0042	0.3508 ± 0.0113
	Deep Ensemble	0.2838 ± 0.0040	0.0630 ± 0.0010	0.0361 ± 0.0003	-0.0254 $\pm$ 0.0005	0.3484 ± 0.0033
Ambiguous	MAP	0.6175 ± 0.0254	0.1249 ± 0.0059	0.0634 ± 0.0021	-0.1559 $\pm$ 0.0031	0.2562 ± 0.0132
	SWAG-Diag	0.6311 ± 0.0271	0.1244 ± 0.0070	0.0632 ± 0.0024	-0.1608 $\pm$ 0.0028	0.2541 ± 0.0090
	MC Dropout	0.4851 ± 0.0082	0.1113 ± 0.0044	0.0598 ± 0.0017	-0.0987 $\pm$ 0.0043	0.3011 ± 0.0113
	TRL (ours)	0.4628 ± 0.0133	0.1098 ± 0.0047	0.0618 ± 0.0018	-0.0351 $\pm$ 0.0037	0.3209 ± 0.0118
	Deep Ensemble	0.4622 ± 0.0060	0.0923 ± 0.0014	0.0540 ± 0.0005	-0.0943 $\pm$ 0.0006	0.3165 ± 0.0049

Deep Ensembles provide the strongest overall match to the human label distribution under KL, soft Brier, and JS on the full set. TRL is second in mean KL and has the highest mean entropy correlation, but its larger soft-target Brier and JS rule out a claim of uniformly better human-distribution matching. On the human-ambiguous subset, TRL nearly matches the five-member ensemble in mean KL (0.4628 versus 0.4622), yields a comparable mean entropy correlation (0.3209 versus 0.3165), and substantially reduces systematic underestimation of average human ambiguity (entropy bias -0.0351 versus -0.0943 for the ensemble and -0.0987 for MC Dropout). Deep Ensembles nevertheless remain clearly stronger under soft Brier and JS. CIFAR-10H thus serves as a complementary diagnostic: TRL tracks relative variation in human perceptual ambiguity and its average level relatively well from a single trained MAP solution, but does not reproduce the full human label distribution as closely as cross-training ensemble diversity.

G Ablations

Unless stated otherwise, the CIFAR-100/ResNet-18 ablations use three posterior-sampling replicates at one fixed MAP checkpoint and the reported default $k_\perp = 30$, $T = 40$, $\beta_\perp = 4.0$, $S = 25$, with 25 rolling-mode FixBN batches. The prior-boost study instead aggregates three independently trained checkpoints. Each sweep is a separate stochastic evaluation, so repeated default rows and runtime values should be compared within, not across, tables. The released code defaults to reset FixBN as described in Appendix C.

G.1 TRANSVERSE RANK $k_\perp$

Table 17: Transverse-rank ablation on CIFAR-100. Mean ± standard deviation over three posterior-sampling replicates at one fixed MAP checkpoint.

$k_\perp$	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑	Runtime (s) ↓
5	0.743 ± 0.002	1.004 ± 0.004	0.076 ± 0.004	0.364 ± 0.002	0.864 ± 0.016	335.585 ± 13.889
10	0.743 ± 0.001	0.983 ± 0.005	0.058 ± 0.006	0.359 ± 0.002	0.865 ± 0.016	401.157 ± 3.204
20	0.744 ± 0.002	0.963 ± 0.006	0.031 ± 0.006	0.357 ± 0.001	0.868 ± 0.018	560.737 ± 19.745
30	0.743 ± 0.003	0.954 ± 0.005	0.0147 ± 0.0003	0.356 ± 0.001	0.870 ± 0.022	706.998 ± 38.901
50	0.741 ± 0.001	0.959 ± 0.004	0.032 ± 0.002	0.360 ± 0.001	0.862 ± 0.017	1036.529 ± 59.444

Rank materially affects predictive performance. With $k_\perp = 5$, NLL and ECE remain substantially above the default configuration. Increasing the rank to 20 captures most of the likelihood improvement, and $k_\perp = 30$ gives the best calibration and Brier score. Moving to $k_\perp = 50$ does not improve the predictive metrics and substantially increases runtime, suggesting that the useful stiff subspace is low-rank but not extremely small. We therefore use $k_\perp = 30$ as a fidelity–cost compromise.

G.2 NUMBER OF STORED SPINE ANCHORS T

Table 18: Stored-spine-anchor ablation on CIFAR-100. Mean ± standard deviation over three posterior-sampling replicates at one fixed MAP checkpoint.

T	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑	Runtime (s) ↓
5	0.743 ± 0.004	0.956 ± 0.005	0.015 ± 0.004	0.356 ± 0.001	0.861 ± 0.020	417.377 ± 5.312
10	0.743 ± 0.002	0.955 ± 0.004	0.014 ± 0.003	0.356 ± 0.001	0.870 ± 0.019	469.773 ± 21.452
20	0.743 ± 0.003	0.955 ± 0.004	0.014 ± 0.004	0.356 ± 0.001	0.869 ± 0.021	538.676 ± 22.166
40	0.743 ± 0.003	0.955 ± 0.006	0.015 ± 0.002	0.356 ± 0.001	0.870 ± 0.022	694.303 ± 40.407

Table 18 varies the number of stored anchors T. Because the CIFAR setting keeps $\Delta s = 0.01$ fixed, this also changes the nominal longitudinal extent $(T-1)\Delta s$; the sweep therefore probes anchor count and extent jointly; it does not isolate discretization density. Performance changes very little from $T = 10$ to $T = 40$, and even $T = 5$ remains competitive. Boldface identifies the best unrounded mean separately for each metric; because the differences among $T = 10$, 20, and 40 are small relative to run-to-run variation and often disappear at the reported precision, the scattered bold entries do not indicate a uniquely preferred anchor count. Thus extending the retained spine beyond $T = 10$–20 adds little clean-test predictive benefit in this setting. The choice $T = 40$ is therefore conservative: it increases runtime but does not materially change clean-test predictive quality relative to $T = 10$–20.

G.3 TUBE SCALE $\beta_\perp$

Table 19: Tube-scale sensitivity on CIFAR-100. Mean ± standard deviation over three posterior-sampling replicates at one fixed MAP checkpoint.

$\beta_\perp$	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑
2.0	0.744 ± 0.002	0.988 ± 0.007	0.067 ± 0.003	0.361 ± 0.002	0.861 ± 0.020
3.0	0.744 ± 0.002	0.965 ± 0.007	0.041 ± 0.004	0.356 ± 0.001	0.865 ± 0.021
4.0	0.743 ± 0.003	0.955 ± 0.006	0.014 ± 0.002	0.356 ± 0.001	0.870 ± 0.022
6.0	0.738 ± 0.004	0.989 ± 0.005	0.063 ± 0.009	0.370 ± 0.001	0.876 ± 0.022
10.0	0.703 ± 0.004	1.257 ± 0.015	0.206 ± 0.014	0.464 ± 0.006	0.862 ± 0.021
20.0	0.387 ± 0.022	2.551 ± 0.056	0.195 ± 0.016	0.823 ± 0.012	0.581 ± 0.095

Table 19 shows how tube scale controls the posterior width in the transported transverse coordinates. In the implemented sampler it is best interpreted as a validation-selected calibration or tempering parameter for the approximate tube: the continuous idealization motivates a transverse covariance

structure, while $\beta_\perp$ sets the effective predictive spread after low-rank truncation and stochastic curvature estimation. The sensitivity curve has a clear optimum for probabilistic prediction: $\beta_\perp = 4.0$ gives the best NLL and ECE and the lowest unrounded Brier score, although Brier rounds to a tie with $\beta_\perp = 3.0$ at three decimals. The $\beta_\perp = 2.0$ and $\beta_\perp = 3.0$ settings tie at the displayed precision in accuracy, while both are worse calibrated than $\beta_\perp = 4.0$. Larger values over-disperse the posterior: at $\beta_\perp = 10$ and 20, samples move into regions that substantially degrade both accuracy and likelihood. The predictive-entropy AUROC is relatively flat for moderate tube scales and peaks at $\beta_\perp = 6.0$ in this rerun, again showing that the scale that optimizes OOD ranking need not optimize calibrated in-distribution prediction. We select $\beta_\perp = 4.0$ because the main objective is probabilistic predictive quality.

G.4 FixBN batches

Table 20: Rolling-mode FixBN batch-count ablation for TRL. Mean $\pm$ standard deviation over three posterior-sampling replicates at one fixed MAP checkpoint.

FixBN batches	Acc ↑	NLL ↓	ECE ↓	Brier ↓	AUROC ↑	Runtime (s) ↓
1	0.741 ± 0.001	1.001 ± 0.009	0.084 ± 0.008	0.372 ± 0.002	0.841 ± 0.025	603.361 ± 10.327
5	0.743 ± 0.003	0.965 ± 0.005	0.043 ± 0.007	0.360 ± 0.001	0.853 ± 0.019	618.030 ± 28.308
10	0.744 ± 0.002	0.955 ± 0.004	0.020 ± 0.007	0.3569 ± 0.0003	0.861 ± 0.020	636.221 ± 28.379
25	0.743 ± 0.003	0.955 ± 0.005	0.014 ± 0.002	0.356 ± 0.001	0.870 ± 0.022	691.244 ± 23.798

Because Table 20 uses the rolling protocol, increasing the batch count changes both the recalibration budget and the amount of running-statistics carry-over from the preceding posterior draw. It should therefore be read as a joint rolling-state/batch-count sensitivity study rather than a pure recalibration-budget ablation. Top-1 accuracy changes little; NLL is similar at 10 and 25 batches, while ECE and Brier continue to change. FixBN uses unlabeled clean training inputs and does not update weights, so it is an inference-time consistency correction, not posterior support. Its necessity does not explain TRL's gain: SWAG-Diag uses its own documented FixBN protocol yet remains near MAP in clean ECE, and the paired rolling-versus-reset audit changes SWAG-Diag only negligibly.

G.5 Backbone prior boost

The large-network TRL prior is a two-block, block-isotropic approximation. The classifier head receives the base precision λ_{base} inherited from the last-layer Laplace fit, while non-head parameters receive

$$\lambda_{\text{backbone}} = \max(c\,\lambda_{\text{base}}, 5).$$

The floor value 5 is a minimum non-head precision that prevents the no-boost setting from becoming an effectively unregularized backbone posterior. Across the three checkpoints used in this ablation, λ_{base} is close to 5, ranging from 4.76 to 4.95; therefore $c = 0$ corresponds to a head/backbone prior of comparable strength, while the floor is inactive for the adopted setting $c = 50$.

Table 21 reports two complementary sensitivity checks on three independently trained ResNet-18 MAP checkpoints in the main CIFAR-100 setting. Panel A fixes the validation-selected tube scale $\beta_\perp = 4$ and varies c, measuring NLL and ECE on both the held-out clean validation split and the test split. On the validation split, the standard validation-NLL selection rule returns $c = 50$, which also attains the best validation ECE. Panel B varies c jointly with $\beta_\perp$, testing on the test split whether the two quantities merely form a single effective scale. The selected value $c = 50$ is then used unchanged in the WideResNet-16-4 and VGG-11-BN architecture-sensitivity checks.

Table 21: Backbone-prior boost ablations on CIFAR-100/ResNet-18. Values are mean ± standard deviation over three MAP checkpoints. Panel A reports validation and test NLL/ECE at the validation-selected tube scale. The value $c = 50$ is selected exclusively by validation NLL; test metrics are reported only after selection. Panel B reports the joint test sweep and shows that the boost c and tube scale $\beta_\perp$ do not reduce to a single effective product.

Panel A: boost sweep at validation-selected $\beta_\perp = 4$

c	$\beta_\perp$	$c\beta_\perp$	NLL (val) ↓	NLL (test) ↓	ECE (val) ↓	ECE (test) ↓
0	4	0	2.000±0.058	1.992±0.058	0.280±0.004	0.283±0.006
10	4	40	1.145±0.010	1.135±0.010	0.160±0.003	0.161±0.006
50	4	200	0.964±0.004	0.954±0.004	0.019±0.002	0.015±0.003
100	4	400	0.972±0.007	0.963±0.007	0.041±0.002	0.041±0.004

Panel B: joint boost–tube-scale test sweep

c	$\beta_\perp$	$c\beta_\perp$	Acc ↑	NLL ↓	ECE ↓	Brier ↓
25	2	50	0.7459±0.0007	0.9652±0.0066	0.0444±0.0036	0.3540±0.0027
25	4	100	0.7438±0.0022	0.9814±0.0026	0.0568±0.0082	0.3652±0.0031
25	8	200	0.6820±0.0066	1.4140±0.0135	0.2520±0.0074	0.5176±0.0064
50	2	100	0.7461±0.0002	0.9834±0.0073	0.0647±0.0023	0.3572±0.0026
50	4	200	0.7461±0.0027	0.9541±0.0041	0.0155±0.0028	0.3537±0.0028
50	8	400	0.7285±0.0033	1.1186±0.0094	0.1547±0.0097	0.4138±0.0044
100	2	200	0.7452±0.0005	1.0006±0.0075	0.0780±0.0020	0.3607±0.0025
100	4	400	0.7462±0.0009	0.9632±0.0065	0.0410±0.0038	0.3537±0.0028
100	8	800	0.7430±0.0024	0.9885±0.0028	0.0647±0.0087	0.3678±0.0029

The two lowest displayed Brier means in Panel B, at $(c, \beta_\perp) = (50, 4)$ and $(100, 4)$, are indistinguishable at the reported precision; no unique Brier optimum is designated without the corresponding unrounded aggregate.

Panel A shows that the two-block prior magnitude is material for calibrated sampling. Validation NLL selects $c = 50$, which also gives the lowest validation ECE. The no-boost setting $c = 0$ collapses, $c = 10$ remains poorly calibrated, and $c = 100$ over-regularizes, producing a U-shaped response in the backbone-to-head precision ratio. Because $c = 10$ already separates backbone and head precision but still performs poorly, the result points to the magnitude of the boost rather than merely the existence of a two-block distinction. Test metrics, evaluated only after validation selection, preserve the same ordering.

Panel B shows joint sensitivity to the boost and tube scale. Equal values of $c\beta_\perp$ do not imply equivalent behavior: the three configurations with $c\beta_\perp = 200$ have test ECE values 0.2520, 0.0155, and 0.0780. We therefore select c and $\beta_\perp$ separately. This sweep rules out $c\beta_\perp$ itself as a sufficient one-dimensional summary; it does not identify a unique effective scalar scale for the joint behavior.

H Transported Eigenspace Diagnostics

Source of variability. Unless a caption states otherwise, seed counts in this appendix index checkpoint-specific runs; they do not denote repeated posterior draws from one fixed checkpoint. The low-loss and functional-disagreement studies use five independently trained MAP checkpoints and their corresponding TRL constructions. The more expensive stale-eigenspace, basis-transport, single-checkpoint, and fresh-refresh studies use the three independently trained checkpoint-specific runs for seeds 0–2.

This appendix measures the consequences of transporting the transverse eigenspace instead of recomputing it locally. We first verify that the spine remains in a low-loss region and quantify its functional drift. We then measure eigenspace rotation and test its predictive effect through comparisons with a fixed MAP basis, a single-checkpoint tube, and a fresh-refresh variant.

The diagnostics show substantial rotation of the local stiff subspace, but the predictive behavior in this benchmark is captured mainly by the MAP-identified low-rank transverse subspace and tube-scale calibration. FixBN supplies the BatchNorm consistency correction required for sampled-weight prediction. The spine remains in a low-loss corridor with limited but nonzero functional variation, and its incremental contribution on clean CIFAR-100 is modest.

The experiments below also quantify the fidelity–cost trade-off behind the practical choices in Section 4.1. The ideal tube suggests exact low-curvature tangents and local transverse covariance

factors, but for the main ResNet/CIFAR-100 regime, the predictive payoff of making these objects more exact is limited compared with the cost. The important mechanism is not exact tracking of a continuous geodesic; it is the combination of a MAP-identified stiff subspace, a low-loss complement trajectory, validation-calibrated transverse scale, and posterior averaging, together with the BatchNorm recalibration needed for consistent sampled-network evaluation.

H.1 Low-loss spine diagnostic

The tubular construction assumes that the discrete spine anchors remain in the same low-loss basin as the MAP solution. We verify this assumption directly by evaluating the validation cross-entropy at every spine anchor after applying the same FixBN recalibration protocol used for posterior prediction. Let $\ell_{\mathrm{val}}(\theta)$ denote validation mean cross-entropy after FixBN recalibration. For a spine anchor γ_t, we report the absolute loss change

$$\Delta_{\mathrm{loss}}(t) = \ell_{\mathrm{val}}(\gamma_t) - \ell_{\mathrm{val}}(\gamma_0), \tag{22}$$

In Eq. (22), γ_0 denotes the MAP anchor of the discrete spine. This diagnostic is separate from hyperparameter selection: it is not used to choose $\beta_\perp$, rank, or stored-anchor count. Its purpose is to check whether the predictor–corrector trajectory remains within a low-loss corridor without relying on posterior averaging to recover from high-loss anchors.

Table 22: Low-loss spine diagnostic on CIFAR-100 with ResNet-18. We evaluate validation cross-entropy at all discrete spine anchors after FixBN recalibration and report changes relative to γ_0. Values are mean $\pm$ standard deviation over five independently trained checkpoint-specific runs, one per MAP checkpoint.

Quantity	Value
Mean Δ_{loss} along spine	0.0030 ± 0.0031
Mean $\lvert\Delta_{\mathrm{loss}}\rvert$ along spine	0.0045 ± 0.0015
Maximum Δ_{loss} along spine	0.0131 ± 0.0039
Endpoint Δ_{loss}	0.0098 ± 0.0040
Validation CE at γ_0	1.0298 ± 0.0384
Validation CE at endpoint	1.0396 ± 0.0410

Across the five checkpoint-specific runs in Table 22, validation loss remains close to the MAP anchor throughout the constructed path. The mean loss change along the spine is 0.0030 ± 0.0031, and the endpoint increase is 0.0098 ± 0.0040 cross-entropy units relative to a base validation loss of 1.0298 ± 0.0384. Some intermediate points decrease the validation loss for individual seeds, and the maximum increase along each spine remains small. After BatchNorm recalibration, the stored trajectory therefore stays within a narrow low-loss corridor.

H.2 Functional disagreement along the spine

Low validation loss alone does not determine whether the spine changes the represented predictor or merely follows near-exact functional invariances. We therefore compare predictions at each spine anchor γ_t with the MAP anchor γ_0 on the validation split after the same FixBN recalibration used for posterior prediction. We report top-1 disagreement, Jensen–Shannon (JS) divergence, cosine distance between predictive probability vectors, and the corresponding validation cross-entropy change, denoted $\Delta_{\mathrm{CE}}(t) = \ell_{\mathrm{val}}(\gamma_t) - \ell_{\mathrm{val}}(\gamma_0)$ within this diagnostic run.

Table 23: Functional disagreement along the CIFAR-100 ResNet-18 spine. We compare predictions at each spine anchor γ_t with the MAP anchor γ_0 after FixBN recalibration. Values are mean $\pm$ standard deviation over five independently trained checkpoint-specific runs, one per MAP checkpoint.

Quantity	Value
Mean top-1 disagreement along spine	0.0316 ± 0.0028
Maximum top-1 disagreement along spine	0.0523 ± 0.0050
Endpoint top-1 disagreement	0.0508 ± 0.0054
Mean JS divergence along spine	0.0028 ± 0.0004
Endpoint JS divergence	0.0063 ± 0.0009
Mean cosine distance along spine	0.0053 ± 0.0008
Endpoint cosine distance	0.0117 ± 0.0017
Endpoint Δ_{CE}	0.0104 ± 0.0028

At the endpoint in Table 23, about 5.1% of validation examples change top-1 prediction relative to γ_0, with JS divergence 0.0063 ± 0.0009 and a validation cross-entropy increase of 0.0104 ± 0.0028. The spine is therefore not functionally identical to the MAP anchor, even though it remains low-loss. The low-loss profile in Table 22 and the functional-disagreement summary here come from separate five-checkpoint diagnostic executions, each applying FixBN within its own run. Their endpoint cross-entropy aggregates differ slightly for that reason; both place the endpoint approximately 0.01 cross-entropy units above γ_0.

Although this longitudinal variation is measurable, averaging over it does not measurably improve predictions beyond the single-checkpoint sampler in Table 26. At this magnitude, the induced predictive change is comparable to seed-level variability and is largely redundant with local transverse averaging. On CIFAR-100/ResNet-18, most of TRL's predictive gain is already captured by the curvature-selected transverse subspace identified at the MAP; the spine supplies low-loss longitudinal support but less additional predictive diversity in this regime.

H.3 Subspace similarity and eigenvalue drift

The practical TRL implementation transports a low-rank transverse structure without recomputing a fresh local eigendecomposition at every spine anchor. This section quantifies the geometric drift introduced by that approximation. Let $Q_t^{\mathrm{tr}} \in \mathbb{R}^{K \times k}$ denote the transported transverse basis at spine location t, let Q_t^{fr} denote a freshly recomputed local top-k eigenspace at the same location, and let Q_0 denote the original MAP basis held fixed along the spine. Here Q_t^{tr} is the diagnostic notation for the same transported basis denoted by N_t in the sampler of Eq. (4). For two orthonormal bases $Q_a, Q_b \in \mathbb{R}^{K \times k}$, we measure subspace similarity by

$$\mathcal{O}(Q_a, Q_b) = \frac{1}{k} \|Q_a^\top Q_b\|_F^2, \tag{23}$$

Here $\|\cdot\|_F$ denotes the Frobenius norm. The overlap in Eq. (23) is invariant to rotations and permutations within the two k-dimensional subspaces. This is important because individual eigenvectors inside a nearly degenerate stiff subspace are not identifiable. We report both $\mathcal{O}(Q_t^{\mathrm{tr}}, Q_t^{\mathrm{fr}})$ and $\mathcal{O}(Q_0, Q_t^{\mathrm{fr}})$, together with principal-angle and eigenvalue-drift summaries. The fresh stiff subspace rotates substantially along the spine: in Table 24, transported-fresh overlap falls to approximately 0.20 at the endpoint.

Table 24: Stale-eigenspace mechanism diagnostic along the TRL spine. Mean $\pm$ standard deviation over three independently trained checkpoint-specific runs (training seeds 0–2). Transported-fresh overlap denotes $\mathcal{O}(Q_t^{\mathrm{tr}}, Q_t^{\mathrm{fr}})$, and MAP-fresh overlap denotes $\mathcal{O}(Q_0, Q_t^{\mathrm{fr}})$.

Spine fraction	Transported-fresh overlap ↑	MAP-fresh overlap ↑	Mean angle (deg) ↓	Mean eigenvalue drift ↓
0.00	1.000 ± 0.000	1.000 ± 0.000	0.0 ± 0.0	0.000 ± 0.000
0.26	0.705 ± 0.043	0.710 ± 0.043	31.6 ± 3.3	0.080 ± 0.044
0.51	0.418 ± 0.035	0.422 ± 0.036	53.1 ± 3.0	0.192 ± 0.038
0.74	0.250 ± 0.045	0.252 ± 0.046	66.3 ± 3.8	0.266 ± 0.070
1.00	0.198 ± 0.021	0.198 ± 0.021	70.8 ± 1.8	0.311 ± 0.076

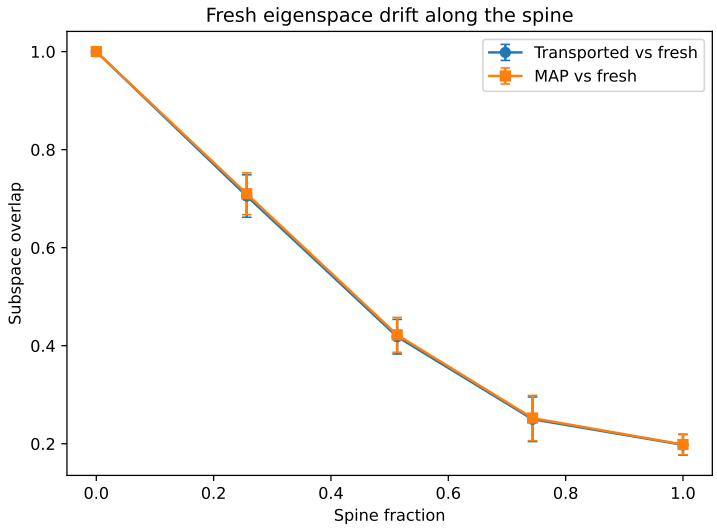


Figure 7: Subspace overlap along the TRL spine. Transported and fixed-MAP bases both drift from the fresh local eigenspace as the spine moves away from the MAP.

Transported–fresh and MAP–fresh overlap are nearly identical throughout the path in Figure 7; the transported basis is no closer to the fresh local eigenspace than the fixed MAP basis. The principal angles and eigenvalue drift likewise show the local stiff subspace rotating away from the MAP. Projection–QR is a transport heuristic rather than a local eigenspace tracker, so the observed gains cannot be attributed to accurate local curvature tracking. Section H.5 tests the more expensive alternative of recomputing Q_t^{fr} at every anchor.

H.4 Predictive effect of transported versus fixed-MAP basis

To test whether the observed eigenspace rotation matters predictively, we compare two samplers on the same constructed spine: one uses the transported basis Q_t^{tr}, and the other reuses the fixed MAP basis Q_0 at all anchors. They are tied within seed variability in Table 25, while both variants substantially improve over deterministic MAP in the main CIFAR-100 results. Transporting the MAP basis provides no measurable predictive advantage over keeping it fixed in this benchmark, so the CIFAR-100 gains do not depend on accurate local curvature tracking. The next subsection tests fresh local recomputation at every spine anchor.

Table 25: Mechanism diagnostic: predictive effect of transported transverse basis versus a fixed MAP basis. Mean ± standard deviation over three independently trained checkpoint-specific runs (training seeds 0–2). No unique winner is highlighted because the variants are tied within checkpoint-level variability.

Variant	Acc ↑	NLL ↓	ECE ↓	Brier ↓
TRL-fixed-MAP-basis	0.743 ± 0.004	0.955 ± 0.004	0.015 ± 0.004	0.356 ± 0.002
TRL-transported-basis	0.745 ± 0.004	0.956 ± 0.004	0.016 ± 0.003	0.355 ± 0.001

H.5 Fresh-refresh and single-checkpoint checks

The preceding ablation compares two inexpensive choices of transverse basis along the same constructed spine. We additionally test two endpoint alternatives. The first removes longitudinal mixing by using only the MAP checkpoint with the same transverse sampling recipe (TRL-single-checkpoint). The second recomputes a fresh local transverse eigenspace Q_t^{fr} at every spine anchor and samples from those refreshed bases (TRL-fresh-refresh-basis). This fresh-refresh variant is not part of the

default TRL algorithm; it is a diagnostic upper-cost variant that tests whether local curvature refresh gives a qualitatively different posterior predictor.

Table 26: Mechanism diagnostic: additional spine and curvature-refresh checks on CIFAR-100 with ResNet-18. Values are mean ± standard deviation over three independently trained checkpoint-specific runs (training seeds 0–2). The fresh-refresh variant recomputes a rank-30 eigenspace at every spine anchor before posterior prediction; the construction time is extra offline cost and is not part of the default TRL pipeline.

Variant	Acc ↑	NLL ↓	ECE ↓	Brier ↓	Extra refresh construction
TRL-single-checkpoint	0.741 ± 0.004	0.958 ± 0.005	0.015 ± 0.003	0.357 ± 0.002	–
TRL-fresh-refresh-basis	0.743 ± 0.003	0.955 ± 0.008	0.013 ± 0.003	0.356 ± 0.002	89.2 ± 7.7 min

The single-checkpoint tube in Table 26 already captures most of the clean CIFAR-100 improvement over deterministic MAP: its NLL and ECE lie in the same range as the full-spine transported and fixed-MAP variants in Table 25. Fresh refresh brings only a small numerical improvement, while recomputing eigenspaces at all 40 stored spine locations, including the MAP anchor, costs 89.2 ± 7.7 minutes per seed as an additional offline step. Together with the modest functional drift in Table 23, these results favor computing the transverse structure once at the MAP for this benchmark. Curvature refresh or longer spines may still be useful in other regimes.

I Small-data Fine-tuning Diagnostic

The main CIFAR-100 diagnostics above show that, for a large network trained from scratch, most of TRL's predictive gain is already captured by the MAP-centered transverse tube. This does not imply that the longitudinal component is always negligible. To test a more favorable non-toy regime, we run an exploratory small-data fine-tuning diagnostic: a ResNet-18-CIFAR checkpoint trained on CIFAR-100 is adapted to CIFAR-10 using 100 training examples per class and a disjoint 100-example-per-class validation split. The official CIFAR-10 test set is used only for final reporting. The configuration is selected as an existence diagnostic and is not a benchmark protocol: we fine-tune for 50 epochs with learning rate 0.003, then run TRL with $T = 20$ stored anchors, $k_\perp = 30$, tube scale 0.1, prior base 5, non-head boost $c = 50$, 25 posterior samples, and 25 FixBN batches per draw. The diagnostic fixes $\lambda_{\text{base}} = 5$, $c = 50$, and $\beta_\perp = 0.1$ as operating values; unlike the main CIFAR benchmark, they are not inherited from its marginal-likelihood fit or clean-validation tube-scale selection. Because the target dataset, fine-tuned checkpoint, active curvature, and projected prior differ, these numerical scales should not be compared directly with the main CIFAR values. Head precision is 5 and non-head precision is $\max(50 \cdot 5, 5) = 250$, so the minimum-precision floor is inactive. Final test evaluation and functional-variance estimation for every stochastic posterior variant use $S = 25$. In the full-spine variant, each draw samples one of all 20 stored anchors uniformly and then samples a transverse perturbation; in fixed-center variants, the anchor index is fixed.

We compare deterministic MAP fine-tuning, a MAP-centered single-checkpoint tube, the full uniform spine mixture, the endpoint-centered tube, and a validation-selected single spine anchor. The last control is a targeted center-selection diagnostic, not an exhaustive best-anchor oracle: validation selection evaluates a six-anchor subset of the 20 stored anchors using $S = 8$ predictive samples per candidate, after which the selected anchor is frozen and reevaluated on the test set with $S = 25$. The restricted candidate set and smaller selection-time Monte Carlo budget make this a weaker control than an exhaustive 20-anchor search. Even so, it separates two effects: whether the spine finds a better local center and whether uniform averaging over all stored anchors adds useful predictive dispersion.

Table 27: Small-data CIFAR-100 → CIFAR-10 fine-tuning diagnostic. A ResNet-18-CIFAR checkpoint trained on CIFAR-100 is fine-tuned on 100 CIFAR-10 examples per class. Values are mean ± standard deviation over ten seeds. The single-anchor control selects from six of the 20 stored anchors on a disjoint validation subset using $S = 8$, then freezes the selected anchor and evaluates it on the official test set using $S = 25$; all reported stochastic metrics use $S = 25$. Avg. func. var. measures variance across posterior predictive class-probability functions; it is a dispersion diagnostic for which larger values are not intrinsically better, and it is not applicable to deterministic MAP.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓	Avg. func. var.
MAP-finetuned	0.777 ± 0.005	0.649 ± 0.014	0.018 ± 0.003	0.315 ± 0.005	–
TRL-single-checkpoint	0.776 ± 0.005	0.649 ± 0.014	0.019 ± 0.003	0.315 ± 0.005	0.00078 ± 0.00010
TRL-best-single-val	0.777 ± 0.005	0.647 ± 0.015	0.019 ± 0.004	0.314 ± 0.006	0.00094 ± 0.00008
TRL-endpoint-single	0.778 ± 0.005	0.647 ± 0.014	0.019 ± 0.004	0.314 ± 0.006	0.00096 ± 0.00011
TRL-full-spine	0.778 ± 0.005	0.647 ± 0.015	0.017 ± 0.003	0.314 ± 0.006	0.00230 ± 0.00043

Table 27 shows that this fine-tuning regime is qualitatively different from the from-scratch CIFAR-100 benchmark. Relative to the MAP-centered tube, full-spine mixing improves NLL and Brier in all ten seeds, ECE in eight, and average posterior functional variance in all ten. It does not consistently outperform the validation-selected anchor; seed 6 is a counterexample. The NLL and Brier gains are small but consistently signed. Posterior functional variance rises from 7.76×10^{-4} to 2.30×10^{-3} on average, about a three-fold increase, showing that longitudinal mixing injects measurable functional variation in this small-data adaptation setting.

The single-anchor controls indicate that center relocation explains most of the NLL/Brier change: the endpoint and validation-selected tubes nearly match the full-spine mixture. Against the validation-selected single-anchor control, full-spine has lower NLL in seven of ten seeds and lower ECE in eight of ten, while Brier is essentially tied. The endpoint requires no search but is fixed by path length rather than by a predictive criterion, while validation selection requires an additional anchor search. Uniform mixing avoids either discrete choice and increases average functional variance to 2.30×10^{-3}, compared with roughly 0.9×10^{-3} for the single-anchor controls, while retaining comparable NLL and Brier. Thus, in this regime, relocation explains most of the mean-score change, whereas mixing primarily preserves longitudinal predictive diversity. The contrast with from-scratch CIFAR-100 is clear: the MAP-identified transverse subspace dominates there, while small-data fine-tuning leaves stronger functional variation along the spine.

For this diagnostic, top-1 disagreement is the fraction of validation examples whose predicted class differs from the initial fine-tuned point, JS divergence is the Jensen–Shannon divergence between predictive probability vectors, and Δ_{CE} is the validation cross-entropy change relative to the initial point. *Mean along spine* averages this quantity over discrete spine anchors, whereas *endpoint* reports the final spine anchor.

Table 28: Longitudinal spine-signal diagnostic in the CIFAR-10 small-data fine-tuning regime. We compare deterministic predictions at discrete spine anchors with the initial fine-tuned point on the disjoint validation split, without transverse posterior samples. Values are mean ± standard deviation over ten seeds.

Quantity	Value
Mean top-1 disagreement along spine	0.0519 ± 0.0104
Endpoint top-1 disagreement	0.0995 ± 0.0180
Mean JS divergence along spine	0.0065 ± 0.0015
Endpoint JS divergence	0.0175 ± 0.0036
Mean Δ_{CE} along spine	0.0191 ± 0.0087
Endpoint Δ_{CE}	0.0594 ± 0.0239

Figure 2 in the main text summarizes this regime contrast. Tables 23, 27, and 28 provide the underlying values and separate the functional-drift diagnostic from posterior predictive performance.

We next ask whether the spine geometry is itself informative about when longitudinal mixing should matter. Across the two evaluated regimes, the functional-disagreement diagnostic from Appendix H

co-varies with the additional dispersion induced by full-spine mixing. We use it as a regime-level characterization of effective longitudinal variation, not as a scalar predictor of the performance gain from mixing. On CIFAR-100 from scratch (Table 23), the transported spine stays functionally close to the MAP: at the endpoint it changes only $5.08\% \pm 0.54\%$ of top-1 predictions, with endpoint JS 0.0063 ± 0.0009 and endpoint Δ_{CE} 0.0104 ± 0.0028 over five seeds. In the CIFAR-10 small-data fine-tuning regime (Table 28), the corresponding endpoint signal is markedly larger: $9.95\% \pm 1.80\%$ top-1 disagreement, endpoint JS 0.0175 ± 0.0036, and endpoint Δ_{CE} 0.0594 ± 0.0239 over ten seeds. The same ordering holds when averaged over the spine: mean JS rises from 0.0028 ± 0.0004 to 0.0065 ± 0.0015. In the measured seeds, the seed-level ranges are separated on endpoint disagreement, endpoint JS, and endpoint Δ_{CE}; endpoint disagreement and JS are therefore roughly $2\times$ and $2.8\times$ larger under fine-tuning. In this terminology, the from-scratch CIFAR-100 corridor has nearly null effective longitudinal variation for prediction, whereas the fine-tuning corridor has a moderate one. This is also the regime in which full-spine mixing increases posterior functional variance and improves over the single-checkpoint posterior; on CIFAR-100 from scratch, the weak longitudinal signal is consistent with the near-redundancy of full-spine mixing and local transverse sampling.

The larger longitudinal signal does not by itself make the spine better: the fine-tuning trajectory also drifts more in validation loss. These are separate questions—whether the path stays in a low-loss corridor and whether movement along it changes predictions—so we report both quantities.

As an auxiliary diagnostic, we repeated the spine-signal measurement on the same $n = 100$ fine-tuning split while varying the trainable subspace (three seeds). Endpoint JS divergence is near zero for head-only fine-tuning (0.00020 ± 0.00004), modest when the last residual block and head are trained (0.0045 ± 0.0003, despite freeing $\approx 75\%$ of the parameters), and largest under full fine-tuning on the same three-seed subset (0.0198 ± 0.0028; the corresponding ten-seed estimate is reported in Table 28). The longitudinal signal is not explained by trainable-parameter count alone; it depends on which parts of the representation are allowed to move. This auxiliary sweep characterizes spine geometry; predictive performance is not evaluated. Endpoint Δ_{CE} rises together with JS across the three modes (-0.0045, 0.0013, and 0.0685), so the stronger longitudinal signal also comes with larger validation-loss drift. Because we did not run the full posterior comparison of Table 27 in the restricted-subspace modes, their predictive effect remains unresolved.

J IMAGENET SCALE-CHECK

To test scalability, we evaluate a pretrained ResNet-50 on ImageNet under three questions: whether the transverse tube calibrates a large model, whether the spine adds functional diversity, and whether rank $k_\perp = 30$ remains sufficient. We treat this experiment as a scale check rather than a standalone benchmark.

Protocol. We use the torchvision ResNet-50 `IMAGENET1K_V1` checkpoint as a fixed pretrained deterministic predictor, not as an exact MAP solution under TRL's two-block prior. Because the checkpoint was pretrained on the official ImageNet training set, a split carved from that set would not be out-of-sample with respect to the checkpoint. We follow the standard post-hoc calibration convention and split the official ImageNet validation set into a 25k tuning half and a 25k held-out evaluation half (Guo et al., 2017). Hyperparameters are selected by NLL on the tuning half and reported only on the held-out half. The MAP gate gives full-validation accuracy 0.7615, matching the canonical torchvision accuracy 0.7613. The MAP accuracy 0.7611 in Table 32 is measured on the held-out half used by the paired rank diagnostic and is therefore not directly comparable to those full-validation values.

Unless otherwise stated, we use $S = 25$ posterior predictive samples, FixBN with 25 training batches per sample, a rank-30 transverse subspace, and the same two-block prior used in the CIFAR experiments: the classifier head uses the fitted base precision λ_{base}, while the backbone uses $\max(c\lambda_{\mathrm{base}}, 5)$. For seed 0, $\lambda_{\mathrm{base}} = 12.5651$. The rank-30 basis is obtained by Lanczos with 90 iterations and 5 HVP batches; its leading-to-trailing Ritz spectrum is 348.6 to 96.2. For the single-checkpoint experiment in Table 29, hyperparameter selection uses the grid $c \in \{50, 150, 450\}$ and $\beta_\perp \in \{0.5, 1, 1.5, 2, 3, 4\}$. The rank ablation below fixes $c = 50$ and sweeps $\beta_\perp \in \{2, 3, 4\}$. This narrower grid is anchored at the selected rank-30 operating point: for seed 0, the broader single-checkpoint grid already selected

$\beta_\perp = 2$, so the rank sweep tests whether adding more transverse directions helps at the calibrated scale without a separate low-temperature re-optimization for each rank.

Single-checkpoint transverse tube. Table 29 reports the single-checkpoint transverse variant on ten random splits, using only the tube around the fixed checkpoint and no longitudinal spine. TRL reduces ECE in all ten held-out halves of the official validation set, while NLL and accuracy incur small, consistent costs.

Table 29: ImageNet/ResNet-50 single-checkpoint transverse TRL over ten random splits. Hyperparameters are selected by NLL on the tuning half of the official validation set and reported on the held-out half. $\Delta = \mathrm{TRL} - \mathrm{checkpoint}$. Seeds 0–2 are the original scale-check; seeds 3–9 are the extension under the same grid and protocol. Relative ECE reduction is computed per split as $(\mathrm{ECE}_{\mathrm{ckpt}} - \mathrm{ECE}_{\mathrm{TRL}})/\mathrm{ECE}_{\mathrm{ckpt}}$ and then averaged over the ten splits; the reported $32.7\% \pm 9.8\%$ is the mean $\pm$ sample standard deviation of those per-split ratios, computed from unrounded per-split ECE values, not a ratio of averaged ECE values.

Seed	Selected $(c, \beta_\perp)$	TRL ECE	Checkpoint ECE	ΔECE	ΔNLL	ΔAcc
0	$(50, 2)$	0.0207	0.0383	-0.0176	$+0.0066$	-0.0043
1	$(50, 2)$	0.0218	0.0377	-0.0159	$+0.0070$	-0.0040
2	$(50, 1.5)$	0.0258	0.0361	-0.0102	$+0.0011$	-0.0007
3	$(150, 3)$	0.0252	0.0408	-0.0156	$+0.0107$	-0.0028
4	$(450, 4)$	0.0250	0.0382	-0.0132	$+0.0050$	-0.0044
5	$(50, 1.5)$	0.0303	0.0370	-0.0067	$+0.0074$	-0.0028
6	$(150, 2)$	0.0287	0.0367	-0.0080	$+0.0050$	-0.0008
7	$(150, 3)$	0.0234	0.0361	-0.0127	$+0.0027$	-0.0026
8	$(150, 2)$	0.0316	0.0400	-0.0084	$+0.0060$	-0.0013
9	$(150, 3)$	0.0210	0.0359	-0.0148	$+0.0055$	-0.0043
Mean	–	0.0254	0.0377	-0.0123	$+0.0057$	-0.0028
Std.	–	0.0038	0.0017	0.0038	0.0026	0.0015

The mean relative ECE reduction, computed per split and then averaged, is $32.7\% \pm 9.8\%$ (mean $\pm$ sample standard deviation across the ten per-split ratios). Because NLL and accuracy are slightly worse in every aggregate comparison, the pattern is a calibration–accuracy trade-off rather than a global predictive improvement. These splits quantify sensitivity to the calibration/evaluation partition; they are not independent model or dataset replicates. The selected configurations remain on the low-NLL plateau of the tuning grid; the larger values of c selected in several extension splits reflect split-specific posterior scaling, with no change to the structural protocol.

Temperature-scaling control. We additionally fit one scalar temperature per split by minimizing NLL on the same tuning half and evaluate it once on the held-out half. Table 30 reports the ten-split result. Temperature scaling preserves the checkpoint argmax and improves all three probabilistic metrics: NLL falls by 0.0131, ECE by 0.0164, and Brier by 0.0011. Its mean temperature $\tau = 1.1297 > 1$ indicates moderate checkpoint overconfidence.

Table 30: ImageNet/ResNet-50 temperature scaling on the same ten random tuning/evaluation splits. Temperature is selected by tuning-half NLL and frozen before held-out evaluation.

Method	Acc ↑	NLL ↓	ECE ↓	Brier ↓	Temperature
Checkpoint	0.760592 ± 0.001540	0.964229 ± 0.006882	0.037676 ± 0.001690	0.336113 ± 0.001582	–
Checkpoint+TS	0.760592 ± 0.001540	0.951139 ± 0.006371	0.021269 ± 0.001705	0.335003 ± 0.001472	1.129749 ± 0.004098

MAP+TS is the stronger scalar calibrator in this pretrained-checkpoint regime: it preserves accuracy, improves NLL, and attains lower mean ECE than TRL (0.0213 versus 0.0254). The outputs are nevertheless different. Temperature scaling returns one deterministic rescaled predictive distribution; TRL returns parameter-space posterior samples, nonzero functional variance, and the transverse/longitudinal support diagnostics studied below. ImageNet serves here as a scale and mechanism check; when scalar held-out calibration is the only objective, MAP+TS is the better choice.

Seed-count note. The single-checkpoint and temperature-scaling results above use ten random splits. The predictor–corrector-step and rank sweeps below remain seed-0 diagnostics: they are paired dose-response tests within a fixed split and are not new multi-seed benchmarks. We interpret them through their within-run trends while using the ten-split spread in Table 29 as context for the scale of split-to-split variation.

Longitudinal spine dose-response. We next test whether the constructed longitudinal spine contributes additional functional diversity around this pretrained checkpoint. We construct the spine using the same mechanics as in the CIFAR experiments: a combined predictor–corrector step $\Delta s\, v_{\parallel} - \eta_{\mathrm{corr}}(I - v_{\parallel}v_{\parallel}^{\top})\nabla_{\theta}\ell_{\mathcal{B}}$, norm control on the combined displacement, tangent update from the realized displacement, rotating correction batches, and transported transverse bases. This ImageNet diagnostic is parameterized by the number n_{step} of predictor–corrector transitions, not by the CIFAR convention in which T counts stored anchors. We sweep the nominal path budget $n_{\mathrm{step}}\Delta s$ from 0 to 0.48 while keeping $c = 50$, $\beta_{\perp} = 2$, the split, the basis, and $S = 25$ fixed. With every transition retained, $n_{\mathrm{step}} = 0$ yields only the MAP anchor and $n_{\mathrm{step}} > 0$ yields $n_{\mathrm{step}} + 1$ stored anchors.

Across the predictor–corrector sweep in Table 31, the longitudinal component is effectively inert: posterior functional variance, NLL, ECE, and Brier remain flat across step count and nominal path budget and are indistinguishable from the single-checkpoint tube. The sweep spans nominal path lengths from 0 to 0.48, bracketing the CIFAR retained-anchor length $(T-1)\Delta s = 0.39$. The matched-step-size setting $\Delta s = 0.01$ reaches 0.24, whereas the 0.48 endpoint uses $\Delta s = 0.02$; the diagnostic therefore matches the path-length magnitude but not the full CIFAR discretization. The construction is numerically healthy as well: held-out diagnostic loss remains flat along the spine, and the norm-control threshold is not activated.

Table 31: ImageNet/ResNet-50 predictor–corrector-step sweep, seed 0. The tuning half is used with $c = 50$, $\beta_{\perp} = 2$, and $S = 25$. Here n_{step} counts transitions: $n_{\mathrm{step}} = 0$ is the single-checkpoint tube with only the MAP anchor, while retaining every transition gives $n_{\mathrm{step}} + 1$ stored anchors.

Nominal length $n_{\mathrm{step}}\Delta s$	Transitions n_{step}	Δs	Func. var. trace	NLL	ECE	Brier
0.00	0	–	0.01380	0.9635	0.0200	0.3372
0.08	8	0.01	0.01373	0.9635	0.0204	0.3372
0.24	24	0.01	0.01362	0.9638	0.0210	0.3372
0.48	24	0.02	0.01372	0.9647	0.0212	0.3374

The held-out half gives the same conclusion: nominal lengths 0.08, 0.24, and 0.48 yield NLL 0.9705, 0.9709, and 0.9705, respectively, with ECE in the range 0.0195–0.0207. Flatness is assessed within this realization. The spine-free reference is the $n_{\mathrm{step}} = 0$ row of Table 31, evaluated on the tuning half within that table's own realization. We do not compare the held-out sweep against the separately constructed rank-30 record in Table 32. The same near-redundancy of longitudinal mixing appears in CIFAR-100 from-scratch training. In both cases, the deeply trained checkpoint sits in a corridor that behaves mostly as a near-invariance and adds little functional diversity; the fine-tuning regimes of Appendix I show the contrasting case in which nearby anchors are more functionally distinct.

Transverse rank. Finally, Table 32 tests whether the default $k_{\perp} = 30$ is underdimensioned for ImageNet/ResNet-50. We compare ranks 30, 50, and 100 on seed 0, keeping $c = 50$ fixed and selecting $\beta_{\perp} \in \{2, 3, 4\}$ by tuning NLL. Each rank uses a fresh Lanczos basis, with iterations scaled as $3 \times k_{\perp}$. The rank-100 spectrum remains nontrivial, with trailing Ritz value 38.9, so the higher-rank test is not limited by a collapsed Lanczos tail. The rank-30 functional-variance trace (0.0160) is not a duplicate of the $n_{\mathrm{step}} = 0$ value in Table 31 (0.01380): the rank table evaluates the held-out half with a freshly computed Lanczos basis and independent Monte Carlo draws, whereas the spine-length table reports the tuning-half diagnostic with its fixed basis. The numerical difference therefore reflects distinct diagnostic realizations, not inconsistent evaluation of one stored posterior.

Table 32: ImageNet/ResNet-50 transverse rank ablation, seed 0. Hyperparameters are selected by NLL on the tuning half and evaluated on the held-out half. The higher-rank rows report larger functional variance without a material held-out improvement in this split; the rank comparison is not replicated across splits.

Variant	Selected $\beta_\perp$	NLL	ECE	Acc.	Func. var. trace
MAP	–	0.9643	0.0383	0.7611	–
Rank 30	2	0.9710	0.0207	0.7568	0.0160
Rank 50	2	0.9731	0.0183	0.7554	0.0203
Rank 100	2	0.9721	0.0210	0.7552	0.0292

Ranks 50 and 100 do not materially improve the held-out metrics relative to rank 30. Rank 50 lowers ECE by 0.0024 in this seed, but this within-split change is not replicated across splits, while NLL and accuracy worsen. Rank 100 likewise leaves NLL/ECE/accuracy effectively unchanged. The higher-rank rows also report larger functional variance, but this single-split pattern is descriptive and does not establish a general rank–variance law. In this split, adding directions beyond $k_\perp = 30$ brings no robust predictive gain, consistent with the rank saturation observed on CIFAR-100.

Taken together, the ImageNet experiments show that TRL remains operational at ResNet-50 scale: the transverse tube reduces ECE in all ten splits, the longitudinal spine is effectively redundant around this checkpoint, and increasing the transverse rank beyond 30 gives no robust gain. The comparison with temperature scaling also separates posterior functionality from calibration alone. ImageNet therefore serves primarily as a scale and mechanism check; the corresponding storage and construction costs are discussed in Appendix L.

K Full-network support selection versus last-layer and all-weight low-rank Laplace baselines

K.1 What is being compared to last-layer and all-weight low-rank Laplace?

At ResNet scale, all compared posterior approximations make a tractability cut. Last-layer ELA/LLA restrict uncertainty to the classifier; FM-LR-LLA ValScale uses a rank-30 all-weight subspace chosen to preserve a specified linearized predictive covariance; TRL instead uses stochastic HVPs to select a rank-30 full-network transverse support, couples it to a discrete low-loss spine, and evaluates nonlinear sampled networks. The FM comparison is checkpoint-, architecture-, split-, and rank-matched but not prior-matched, so it tests alternative support-allocation strategies; it is not a one-factor subspace ablation. Appendix E.1 gives the full protocol.

K.2 Support-selection diagnostics

TRL uses curvature primarily to choose where low-rank transverse uncertainty is placed, while longitudinal support is restricted to verified spine anchors. The remaining parameter directions stay at their MAP values, so the method is a support-restricted full-network approximation and does not represent a complete posterior over all flat directions. Appendices H and I separately test when longitudinal mixing contributes; the control below asks whether full-network low-rank access alone explains the CIFAR gain.

As an additional full-network control, we evaluated random rank-30 posteriors using the same block prior, the same validation-selected tube-scale grid, $S = 25$ predictive samples, and the same FixBN protocol as TRL. On a fixed MAP checkpoint, five independent random bases produced nearly identical, MAP-like predictions (Acc 0.7413 ± 0.0001, NLL 1.0378 ± 0.0002, ECE 0.0951 ± 0.0001, versus the same checkpoint's MAP values of 0.7399, 1.0498, and 0.0996). One-basis controls on two additional MAP checkpoints showed the same MAP-like behavior, and all seven random controls selected the largest value in the original scale grid, $\beta_\perp = 20$ (overall Acc 0.7418 ± 0.0008, NLL 1.0331 ± 0.0076, ECE 0.0925 ± 0.0041, AUROC 0.8647 ± 0.0128). The predictive-entropy AUROC range overlaps the MAP and TRL ranges in Table 1, consistent with AUROC being less sensitive than NLL/ECE to the calibration mechanism isolated by this control.

Extending one basis to $\beta_\perp = 40$ improved validation NLL only modestly (from 1.016 at $\beta_\perp = 2$ to 1.000 at $\beta_\perp = 40$, a gain of 0.016 across a $20\times$ scale range) and left the predictor MAP-like on test data (Acc 0.742, NLL 1.023, ECE 0.089). In contrast, the curvature-selected TRL subspace reaches its calibrated operating regime at $\beta_\perp = 4$ and collapses at $\beta_\perp = 20$ (Table 19), showing that TRL is sensitive to scale because its selected directions are functionally loaded. Within this matched control, full-network access, low-rank sampling, the block prior, evaluation-consistent FixBN, and validation scale selection do not reproduce the TRL correction; the curvature-selected transverse support is the remaining distinguishing component.

Table 33: Support-selection scale diagnostic. The seven random controls comprise five independent random bases on one fixed MAP checkpoint plus one basis on each of two additional checkpoints; their aggregate is descriptive and mixes basis and checkpoint variation. The curvature-selected TRL row reports three posterior-sampling seeds on one fixed MAP checkpoint, so its uncertainty is not directly comparable to the seven-control aggregate. The representative extension is one random basis evaluated beyond the original scale grid. All variants use the same two-block prior, validation-scale selection, $S = 25$, and FixBN protocol.

Support	Scale protocol	Validation behavior	Test behavior
Random rank-30, seven controls	Original grid $\{2, 3, 4, 6, 10, 20\}$	All seven select the upper boundary, $\beta_\perp = 20$	MAP-like overall: Acc 0.7418 ± 0.0008, NLL 1.0331 ± 0.0076, ECE 0.0925 ± 0.0041
Random rank-30, representative extension	$\beta_\perp : 2 \to 40$	Validation NLL changes only $1.016 \to 1.000$ across a $20\times$ scale range	At $\beta_\perp = 40$: Acc 0.742, NLL 1.023, ECE 0.089
Curvature-selected TRL	Original grid $\{2, 3, 4, 6, 10, 20\}$	Interior optimum at $\beta_\perp = 4$; high scales are rejected	Three sampling seeds on one fixed checkpoint: at 4, Acc 0.743 ± 0.003, NLL 0.955 ± 0.006, ECE 0.014 ± 0.002; at 20, Acc 0.387 ± 0.022, NLL 2.551 ± 0.056, ECE 0.195 ± 0.016

Table 33 makes the qualitative contrast explicit. Random bases remain nearly insensitive to large scale changes and stay MAP-like, whereas curvature-selected support has an interior validation optimum and fails sharply when over-scaled. The control therefore distinguishes the functional loading of the selected support, not merely tuned versus untuned low-rank sampling.

For the five-basis random-subspace control and the seed-0 FM-LR-LLA diagnostic, both pipelines load the same byte-identical CIFAR-100 MAP seed-0 checkpoint, as verified by the release manifest. The comparison is therefore checkpoint-matched. On this checkpoint, MAP gives Acc/NLL/ECE $0.7399/1.0498/0.0996$; the five-basis random rank-30 control gives $0.7413/1.0378/0.0951$ with Brier 0.3694; FM-LR-LLA gives $0.7410/0.9913/0.0463$ with Brier 0.3582; and the seed-0 TRL result reported in Table 1 gives $0.743200/0.959857/0.014815$ with Brier 0.355263. The random support remains close to MAP, whereas the informed low-rank constructions produce substantially larger probabilistic corrections. Within this checkpoint-matched comparison, full-network access and rank 30 are not sufficient to reproduce the TRL correction. It is not, however, a one-factor ablation: FM-LR-LLA uses an all-weight Kronecker Laplace reference, a marginal-likelihood prior, and MC-linearized predictions, whereas the random and TRL controls use TRL's two-block prior, FixBN, and nonlinear network evaluation. Accordingly, the remaining FM–TRL gap should not be attributed solely to subspace selection.

Table 34 summarizes the corresponding diagnostic controls and the failure modes they isolate.

Table 34: Diagnostic controls isolating common failure modes of the practical TRL approximation.

Question	Control
Full-network access?	Random rank-30 full-network bases remain MAP-like; curvature selection matters (Appendix K).
Spine as main CIFAR mechanism?	TRL-single-checkpoint nearly matches full TRL on CIFAR-100, isolating the MAP-identified transverse subspace (Table 26).
Transport fidelity?	Fresh-refresh diagnostics show eigenspace drift, but small predictive gains relative to cost (Appendix H).
Curvature-operator confound?	A seed-0 replacement of the positive-spectrum Hessian selector by exact GGN selection yields nearly identical downstream metrics; this checks predictive robustness, not eigenspace identity (Appendix C.2).
BatchNorm-statistics confound?	The rolling-mode sweep (Table 20, a joint batch-count/carry-over sensitivity) shows that probabilistic metrics change as recalibration budget and carry-over jointly vary; random full-network controls evaluated with the same FixBN protocol remain MAP-like, showing that FixBN alone does not explain TRL's calibration gain (Appendix K).
CIFAR-100-C tuning?	Tube scale is selected on clean validation data and reused unchanged for corrupted-test evaluation.

L Computational Complexity and Sampling Efficiency

Geodesic-based Riemannian Laplace methods and Laplace diffusion generate samples by numerically integrating geometric dynamics. With T_{int} integration steps per draw and per-step cost C_{step}, generating S samples costs $O(S\,T_{\mathrm{int}}\,C_{\mathrm{step}})$. The step count and cost depend on the method, metric representation, and integrator; this expression describes the sampling procedure and is not a measured runtime comparison.

TRL constructs its finite spine and low-rank transverse structure once per MAP checkpoint. Posterior prediction then samples a spine anchor and applies a low-rank map $\theta^{(r)} = \gamma_t + N_t(\beta_\perp L_{\perp,t} z_\perp^{(r)})$, followed by ordinary forward passes and, for BatchNorm networks, FixBN recalibration. Thus TRL trades additional offline storage and construction time for avoiding a fresh geodesic or curvature solve for every posterior sample.

Because the logged timers cover different scopes and the TRL total contains overlapping nested FixBN timers, Table 14 cannot support an end-to-end wall-clock ranking. The robust comparison is computational structure: TRL uses one trained backbone plus post-hoc geometric construction, validation-scale selection, and posterior prediction, whereas the reported Deep Ensemble uses $M = 5$ independently optimized backbones. The demonstrated advantage here is fewer independent optimization runs; runtime superiority over Deep Ensembles is not established.

ImageNet-scale cost. The ImageNet/ResNet-50 scale-check in Appendix J follows the same amortized cost profile at larger scale. The MAP checkpoint is fixed; the additional offline cost comes from HVP/Lanczos construction of the transverse basis and, for spine diagnostics, from storing or staging transported bases. For ResNet-50, a rank-30 transverse basis has roughly 30×25.6M floating-point entries, about 3.1GB in fp32, while the rank-100 ablation increases this to about 10.2GB. At high rank, the Lanczos workspace, not the final basis alone, is the main memory bottleneck, so the rank-100 diagnostic stages the basis on CPU. For full-spine diagnostics, storing a transported basis at every anchor would multiply this cost by the number of anchors; we therefore stage anchor bases on CPU and move the selected basis to GPU only when drawing a posterior sample. Thus the ImageNet experiment preserves the intended TRL trade-off: one offline geometric construction around a fixed MAP checkpoint, followed by ordinary posterior forward passes and FixBN recalibration, with no training of multiple independent models.

M Limitations and Additional Discussion

TRL is a post-hoc, single-MAP approximation and does not capture the cross-basin diversity of Deep Ensembles; its OOD AUROC can therefore be lower than that of methods with more diffuse function-

space uncertainty. It also has offline and prediction-time overhead: spine construction requires HVP/Lanczos operations, and BatchNorm posterior prediction requires FixBN unless an alternative normalization or cached-statistics strategy is used. For architectures without running normalization statistics, including LayerNorm-based Vision Transformers, this FixBN step is unnecessary, although curvature construction and nonlinear posterior-sample evaluation remain. The current implementation uses a one-dimensional spine, a low-rank transverse subspace, and a validation-selected tube scale $\beta_\perp$ as an explicit calibration/tempering parameter. Transported low-rank structure can drift from freshly recomputed local eigenspaces, so exact local geometric fidelity should not be assumed. We leave higher-dimensional spines, periodic curvature refresh, normalization without per-sample recalibration, and cross-basin mixtures to future work.

N Impact Statement

This work studies a post-hoc uncertainty approximation that may support reliability analysis, but it does not establish safety in deployed or high-stakes systems. Improved calibration does not correct dataset bias, distributional mismatch, or failures outside the evaluated settings; deployment therefore requires independent validation and domain-specific safeguards.

O Extended Related Work and Positioning

Riemannian Laplace maps local Gaussians along geodesics (Bergamin et al., 2023; Yu et al., 2024), whereas Laplace diffusion uses metric-dependent random walks (Roy et al., 2024). TRL constructs its posterior approximation in two stages: minibatch gradients guide the predictor–corrector spine, defining its anchors, and posterior samples are then generated by selecting an anchor and drawing a transverse Gaussian perturbation. Reparameterization invariance is not enforced; Appendix H measures local eigenspace drift.

The continuous idealization is related to the geometric-Gaussian design space of Da Costa et al. (2025), but the implemented piecewise spine is not a global smooth diffeomorphic pushforward and inherits no ReparamGA universality guarantee. Function-space methods such as FSP-Laplace (Cinquin et al., 2024) and functional variational BNNs (Sun et al., 2019) instead impose function-space prior structure during training; TRL is post hoc and does not provide their explicit far-from-data prior-reversion guarantees.

Low-rank subspace methods differ mainly in what the retained rank is asked to preserve. Faller & Martin (2026) target the predictive covariance of a specified full linearized GGN–Laplace reference; TRL chooses parameter-space support and evaluates nonlinear sampled networks, with the matched FM-LR-LLA ValScale comparison in Appendix E.1. Izmailov et al. (2020) compare random, SGD-PCA, and curve subspaces; our random rank-30 basis is only a control, while TRL uses HVP-selected transverse support plus discrete anchor mixing. Mode-connectivity methods connect independently trained minima (Garipov et al., 2018; Draxler et al., 2018), whereas TRL starts from one MAP with no prescribed endpoint. Loss-tunnel MCMC (Dold et al., 2025) also targets posterior connectivity but requires a learned tunnel and a prediction-time chain; TRL builds a finite spine with minibatch gradients and then samples from its Gaussian fibers.

P AI Use Statement

Generative AI tools were used for manuscript drafting and language editing, consistency checking, conceptual and methodological feedback, experimental-design feedback, interpretation of experimental results, assistance in implementing research methods and editing/reviewing associated software code, and creation or modification of scientific figures and release artifacts. All AI-assisted text, code, figures, analyses, and artifacts were reviewed and validated by the authors against the manuscript, source code, experiment logs, and released artifacts. The authors take responsibility for the final content of this work, including text, claims, code, figures, and artifacts produced with AI assistance.