MEASURING TRANSFORMATION-COVERAGE SAVINGS FROM GEOMETRIC PRIORS
A FIXED-DATA EVALUATION OF EQUIVARIANCE, AUGMENTATION, AND CAPACITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Geometric priors can reduce the range of transformations needed during training, but measured savings depend on the baseline's training and inference budgets. We evaluate this dependence by fixing image count and test transformations, varying training rotation coverage, and recording the smallest sampled training half-width that reaches a shared accuracy target. A 60-run ImageNet-100 grid and a separate 45-run CIFAR-10 replication both show that exact C_4 saves an average of 156 degrees of training-band half-width on five test angles at a fixed epoch budget. Stronger baselines change the interpretation. On ImageNet-100, a plain model trained for the measured C_4 time and evaluated with four-view averaging reaches 10 of 11 fixed-epoch targets from the augmentation cohort at a half-width of 30 degrees, while C_4 retains accuracy advantages of 4.03 percentage points at that width and 2.97 percentage points at full coverage. On CIFAR-10, the corresponding advantage falls from 14.83 percentage points at 30 degrees to 0.07 percentage points at 90 degrees. Against the stronger full-band target, the primary operator gives 42 degrees of mean per-angle saving but zero saving for the macro-average target; crossings under the alternative operator and at the seed level are partly censored. These results expose sensitivity to the target and sampled grid. Separate learned-prior controls find smaller gains on hierarchical backbones than on flat networks and matched tasks. Coverage savings at fixed epochs therefore need not imply persistent accuracy gains under stronger baselines; transformation support, computation, uncertainty, and implementation fidelity must be reported together.

1 INTRODUCTION

Geometric priors can improve accuracy on transformations poorly represented in training. Their practical value depends on how much training coverage they save, and whether augmentation or additional computation can recover the same benefit. Learned alignment (Jaderberg et al., 2015; Dai et al., 2017; Zhu et al., 2019) and constructed equivariance (Cohen & Welling, 2016; 2017; Weiler & Cesa, 2019) address this problem differently. Comparing their accuracy alone leaves both the required transformation support and the cost of the alternatives unresolved.

We measure transformation coverage as a resource: fix the images and test angle, vary the training rotation band, and record the first sampled band reaching a shared accuracy target. Across five ImageNet-100 test angles, exact C_4 saves an average of 156 degrees of training-band half-width relative to full-band plain at 30 epochs. Folded distance provides a useful, imperfect account: 150° folds to -30° modulo the grid-exact C_4 subgroup. Figure 1 shows how the coverage curves change with budget on a second dataset.

The interpretation changes as the baseline strengthens. On ImageNet-100, exact C_4 uses about four times plain's FLOPs and eight times its training time. Augmentation preserves large fixed-epoch savings, but a plain model trained for the measured C_4 time and evaluated with four-view averaging reaches 10 of 11 reference targets at $T = 30$. C_4 retains 3–4 percentage points with greater inference cost. A complete 45-run CIFAR-10 replication establishes a boundary: narrow-coverage gains

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.

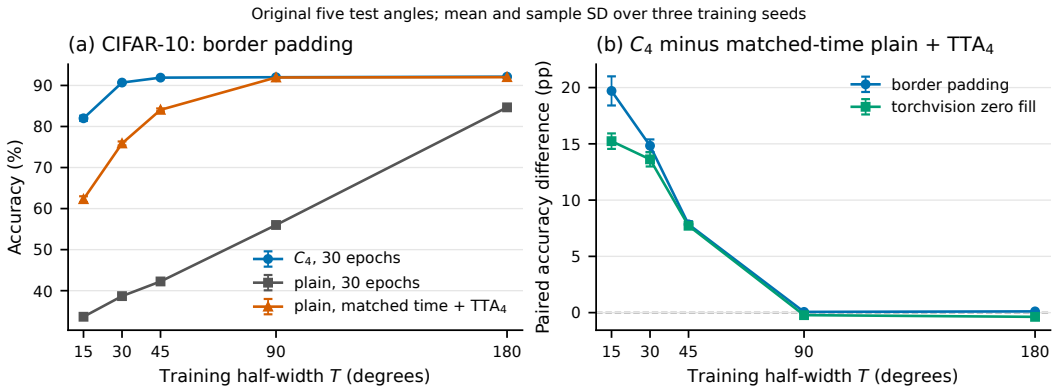


Figure 1: **Coverage advantages depend on the comparison budget.** Full CIFAR-10 replication, original five test angles. **(a)** At fixed epochs, exact C_4 reaches high accuracy with narrow training support; time-matched plain with four-view averaging nearly catches up at larger coverage. **(b)** Paired accuracy gaps under two rotation operators. Points first average angles within each seed; error bars are sample SD over three seeds. These curves show accuracy differences, not equivalence tests.

survive this stronger baseline, whereas at $T = 90$ and 180 the matched-time baseline's accuracy nearly matches that of C_4. Coverage savings at fixed epochs therefore need not imply a persistent accuracy advantage under stronger budgets.

Separate experiments examine a learned orientation prior (GCLS, §4). Its gains persist on flat networks but are smaller on the tested hierarchical backbones, where augmentation performs better. Reducing registration data does not restore the earlier advantage, while widening a flat CNN preserves it. These controls, together with five established mechanisms and matched-task comparisons, are consistent with data quantity alone not explaining the observed pattern; architecture and task remain possible contributors.

Contributions. (1) A fixed-data comparison that measures how much transformation support a prior saves at a shared target, conditional on the baseline and its budget. (2) Coverage replications on two image datasets showing where matched training time and orbit averaging preserve or absorb the accuracy advantage. (3) Generator audits and learned-prior controls that connect each finding to its implementation, architecture, and task.

1.1 Related work

Learned and constructed geometric priors. Spatial transformers, deformable convolution, CoordConv (Liu et al., 2018), and orientation fields encode different geometric biases. ORN (Zhou et al., 2017) also carries constructed orientation structure, so its single-stage insertion and full-network implementation require separate comparisons (§4.4). Group-equivariant and steerable CNNs impose transformation structure by construction, with fidelity depending on discretization. Their use in galaxy morphology (Dieleman et al., 2015; Scaife & Porter, 2021; Pandya et al., 2023) motivates one of our additional endpoint checks.

Equivariance versus augmentation. Brehmer et al. (2024) find that sufficient training with augmentation can close a data-efficiency gap on a rigid-body benchmark, while equivariance remains favorable at the tested compute budgets. Adly (2026) reports that augmentation with orbit averaging matches an equivariant model on a controlled C_n task. We instead hold image count fixed and vary transformation support at fixed test angles. Wider-plain, matched-training-time, and orbit-averaging controls establish how the measured savings depend on the comparator (Appendix Table 24).

Approximate and measured equivariance. Relaxed group convolutions (Wang et al., 2022), residual pathway priors (Finzi et al., 2021), and learned partial equivariance (Romero & Lohit, 2022) soften symmetry constraints. The Lie derivative (Gruver et al., 2023) measures learned equivariance and discretization error. Our generator audit complements coverage curves: the tested C_8 imple-

mentations differ in narrow-band accuracy despite overlapping measured errors. Our contribution is the evaluation and its findings; penalty-based soft equivariance remains outside its scope.

2 PROTOCOL AND COMPARISON CRITERIA

The Geometric-Prior Evaluation Protocol (GPEP) measures coverage savings at fixed image count and checks accuracy gains against stronger alternatives. Its coverage sweeps and learned-prior controls use separate experiments.

What the comparison identifies. Coverage saving is conditional on the target, training budget, and inference procedure. Fixing these comparison rules asks how much angular support each model needs; strengthening the baseline asks whether computation can substitute for that support. We report both the first crossing and the accuracy curve, since a near-plateau difference can move a crossing across a large gap in the sampled grid.

Comparison criteria. We organize the evaluation around four controls:

1. **Augmentation comparison**: compare with augmentation as well as plain. A lower score establishes inferiority to that baseline under the tested conditions; complementarity requires a prior-plus-augmentation arm.
2. **Hierarchical-backbone comparison**: test whether gains on flat networks persist on ResNet, U-Net, ViT, or RAFT, while retaining the task and training details of each comparison.
3. **Constructed-equivariance control**: when testing extrapolation, include an architecture with explicit group structure to contextualize the learned prior's performance.
4. **Generator-level audit**: measure the claimed generator, such as 45° for C_8, and report the input distribution and training checkpoint. A group label or a 90° check alone does not establish C_8 equivariance.

Reporting and uncertainty. Per-seed results accompany mean $\pm$ standard deviation. Paired tests and intervals over three seeds are exploratory and sensitive to small-sample assumptions. A non-significant difference is not an equivalence result. The ImageNet-100 grid has four arms and five coverage levels; the CIFAR-10 replication has three arms at those levels, both with three seeds. Earlier two-seed steerable cells remain separate. Test angles reuse images and models; they are not independent replicates.

Parameter and compute controls. We report parameter counts, including the additional ViT prior parameters. Similar counts control model size but leave architectural and optimization differences. The coverage family includes a wider plain model with approximately $4\times$ FLOPs and plain models trained for the measured time of paired C_4 runs on the same host and GPU. These controls test specific budget allocations.

Coverage and resource reporting. The default target is same-angle, same-operator fixed-epoch plain accuracy at $T = 180$. CIFAR-10 also uses a stronger target: full-band time-matched plain with four-view averaging. We report first sampled crossings without interpolation and retain missing crossings as censored. The 60-run ImageNet-100 grid uses rotation alone; its 18-run augmentation extension and six matched-time controls add translation, flips, and a time-indexed OneCycle schedule. CIFAR-10 covers both budgets at all five bands. Parameters, FLOPs, training time, peak memory, and batched inference time are reported separately; matching training time does not match inference cost.

Transformation support. Hierarchical-backbone comparisons evaluate stronger elastic deformation, with rotation controls for flow and registration. Broad-band tests include both seen and unseen angles; coverage sweeps instead fix each test angle while varying the training band. Generator checks report normalized feature changes at stated angles and checkpoints. All tested C_8 implementations are approximate.

test angle	plain@180 target (%)	d_{C_4} (deg.)	exact C_4 T / saving	interp. C_8 T / saving	steer. C_8 T / saving
105°	73.76	15	15 / 165	15 / 165	15 / 165
120°	73.32	30	30 / 150	30 / 150	15 / 165
135°	73.51	45	30 / 150	30 / 150	15 / 165
150°	73.52	30	30 / 150	30 / 150	15 / 165
165°	73.75	15	15 / 165	15 / 165	15 / 165
mean saving	—	—	**156**	**156**	**165**

Table 1: Measured transformation-coverage saving on the complete ImageNet-100 grid (60 runs, three seeds per cell). The target is the same-angle plain model trained at $T = 180$; each entry reports the smallest sampled T reaching that target, followed by the saved band $180 - T$. The table is descriptive on the measured grid and does not interpolate or claim a universal quotient law. Full accuracy curves, ideal-C_8 references, and resource controls are in the appendix.

3 Measuring transformation-coverage savings

Let $A_a(T, \alpha)$ denote mean accuracy over three seeds for arm a, trained on $[-T, T]$ and evaluated at angle α. We hold image count and test transformation fixed. Unless specified otherwise, the shared target q is the same-angle, same-operator accuracy of fixed-epoch, single-view plain at $T = 180$, within the same augmentation cohort. The first sampled coverage reaching it and the resulting saving are

$$\widehat{T}_a(q; \alpha) = \min\{T \in \mathcal{T} : A_a(T, \alpha) \geq q\}, \qquad \widehat{\Delta T}_a(q; \alpha) = \widehat{T}_{\text{plain}}(q; \alpha) - \widehat{T}_a(q; \alpha), \tag{1}$$

Here T is the training-band half-width: $\widehat{\Delta} T$ measures saved angular support, not fewer examples or continuous sample complexity. Missing crossings are right-censored; the target, split, and test operator are shared across arms. The plain crossing is measured from its curve; it is not automatically set to 180°. This matters when a narrower band already reaches the full-band reference.

3.1 ImageNet-100: coverage savings and resource cost

The ImageNet-100 experiment uses 64px inputs, a fixed 116,689-image training split, and a disjoint 10,000-image evaluation split. Four arms (plain, exact C_4, interpolated C_8, and steerable C_8) train at $T \in \{15, 30, 45, 90, 180\}$ with three seeds each: 60 runs in total. Each run evaluates $105/120/135/150/165^\circ$. Training uses 30 epochs of OneCycle SGD and batch size 512. Table 1 reports mean-based crossings; per-seed crossings and additional gauge, 5×5 steerable, and wide-plain controls appear in the appendix. Intervals reflect seed variability, not model-selection or dataset uncertainty.

Measured savings. Exact C_4 and interpolated C_8 save 150–165 degrees across the five angles, averaging 156 degrees. Steerable C_8 reaches the target at the smallest sampled band, $T = 15$, saving 165 degrees at every angle. The grid cannot resolve smaller crossings. Seed-level crossings and 72%/73% target checks preserve the ordering (Appendix Tables 22 and 23). These savings are conditional on the stated target and fixed-epoch budget.

Cost and alternatives. Exact C_4 uses $4.0\times$ the FLOPs of plain and $8.0\times$ its training time; approximate C_8 uses 7.6–$10.2\times$ the FLOPs and 9.1–$11.4\times$ the time (Table 21). Parameter counts and peak accelerator memory accompany these measurements. A wider plain control (44.8M parameters, approximately $4\times$ the FLOPs of plain) reaches 75.96 ± 0.18 at $T = 180$, below a C_4 model trained at $T = 45$ in the initial 150° comparison. The gauge arm stays close to plain (differences from -1.27 to $+0.38$ percentage points across recorded cells). The stronger augmentation, training-time, and test-time controls below address alternatives that parameter matching alone does not.

3.2 Augmentation, equal training time, and orbit averaging

An 18-run grid repeats plain and exact-C_4 training with translation and horizontal-flip augmentation at $T \in \{15, 30, 180\}$. At 30 epochs, C_4 saves 156 degrees of mean sampled half-width across the original angles and 160 across six additional angles. The six additional angles were selected after

test angles	T	time-matched plain+TTA$_4$	exact C_4	paired difference
original five	30	75.16 ± 0.11	79.19 ± 0.05	$+4.03 \pm 0.16$
original five	180	76.36 ± 0.19	79.33 ± 0.17	$+2.97 \pm 0.33$
additional six	30	75.42 ± 0.08	79.42 ± 0.07	$+4.00 \pm 0.13$
additional six	180	76.30 ± 0.13	79.33 ± 0.24	$+3.03 \pm 0.37$

Table 2: Accuracy under the strongest tested plain baseline. All arms use translation and flip augmentation. Plain receives the paired C_4 training-time budget and four-view probability averaging; C_4 uses one view. Each seed first averages the indicated angles; entries report mean $\pm$ sample SD over three seeds. Differences are C_4 minus plain in percentage points. Training time is matched, but inference cost is not.

T	plain, 30 epochs	C_4, 30 epochs	plain, time+TTA$_4$	paired gap
15	33.65 ± 0.33	81.99 ± 0.64	62.29 ± 0.71	$+19.70 \pm 1.29$
30	38.67 ± 0.46	90.70 ± 0.10	75.86 ± 0.49	$+14.83 \pm 0.56$
45	42.25 ± 0.42	91.90 ± 0.13	84.05 ± 0.34	$+7.85 \pm 0.28$
90	56.01 ± 0.42	92.00 ± 0.06	91.94 ± 0.06	$+0.07 \pm 0.09$
180	84.66 ± 0.23	92.13 ± 0.09	92.02 ± 0.11	$+0.11 \pm 0.02$

Table 3: Full CIFAR-10 replication (50k training and 10k test images). Accuracy averages the original five test angles under border padding within each seed; entries are mean $\pm$ sample SD over three seeds. The strongest plain arm receives the paired C_4 training-time budget at each T and seed, then uses four-view probability averaging. The gap is C_4 single-view minus that arm, in percentage points. Training time is paired within each T, not held globally constant across T; inference cost differs.

observing the training results but fixed before post-evaluation. Four-view probability averaging does not move the fixed-epoch plain crossing below $T = 180$.

Six stronger controls train plain on the same host and GPU for the measured time of each paired C_4 run. With four-view averaging, this baseline remains 4.03 ± 0.16 percentage points below C_4 at $T = 30$ and 2.97 ± 0.33 percentage points below C_4 at $T = 180$ on the original angles (Table 2). The additional angles give similar gaps, and a zero-filled torchvision rotation changes the cohort-mean gaps by less than 0.08 percentage points. At batch 512 on an RTX 4080, the measured inference times per image are 0.470 ms for C_4, 0.348 ms for four-view plain, and 0.087 ms for single-view plain.

The stronger plain baseline reaches 10 of 11 fixed-epoch targets at $T = 30$ under both rotation operators. Thus the 156/160-degree savings describe the fixed-epoch comparison, while the accuracy advantage persists against the stronger baseline. Without matched-time $T = 15$ and intermediate coverage levels, the remaining coverage advantage is unresolved. Appendix C reports all cells, paired-seed variation, and timing details.

3.3 CIFAR-10: A COMPLETE SWEEP WITH MATCHED TRAINING TIME

A separate 45-run experiment uses all 50,000 training and 10,000 test images of CIFAR-10 at 32-pixel resolution. The same ResNet-18 and exact-C_4 families, with ten-class heads, train with rotation, translation, and flip augmentation at all five coverage levels and three seeds. Both fixed-epoch arms train for 30 epochs; a third arm trains plain for the measured training time of each paired C_4 run. The protocol and all eleven nonzero test angles were fixed before these runs, following the ImageNet-100 results. This is a second dataset under synthetic rotations, not a new transformation family.

The fixed-epoch saving is again 156 degrees on the original angles and 160 on the additional angles under both rotation operators. Against time-matched plain with four-view averaging, however, the original-angle C_4 advantage falls from 14.83 ± 0.56 percentage points at $T = 30$ to 0.07 ± 0.09 at $T = 90$ and 0.11 ± 0.02 at $T = 180$ (Table 3). Zero-filled rotation reverses the full-band difference to -0.37 ± 0.11 percentage points. Narrow-coverage gains therefore coexist with nearly equal broad-coverage accuracy; the ImageNet-100 residual gap does not transfer in magnitude.

Using full-band time-matched plain with four-view averaging as the stronger target, original-angle savings are 60, 90, 0, 0, and 60 degrees under the primary operator, averaging 42 degrees. Crossing the mean accuracy across these angles gives a different statistic: both curves first reach the target at $T = 180$, yielding zero saving. At $T = 90$, the mean accuracy misses its target by only 0.014 percentage points, illustrating the sensitivity of a strict crossing near a plateau. One additional angle and three original angles under zero fill have no C_4 crossing. These censored outcomes, seed sensitivity, and costs appear in Appendix D; the 42-degree mean is descriptive and not an operator-independent saving.

3.4 Folded distance and the limits of the subgroup account

Let planar rotations G act on inputs, let $G_T = \{g : |\mathrm{angle}(g)| \leq T\}$ be the training band, and let H be a subgroup under which the classifier is invariant. Define the folded distance and residual coverage gap as

$$d_H(g^*) = \min_{h \in H} d(h^{-1}g^*, e), \qquad u_H(g^*, T) = \max\{d_H(g^*) - T, 0\}, \tag{2}$$

where d is circular angular distance. Exact invariance gives $f(hgx) = f(gx)$, extending predictions from G_T to HG_T. It does not guarantee correct predictions, independence from the training distribution, or generalization beyond this set (Appendix F). The empirical hypothesis is that accuracy recovery with increasing T follows residual folded distance more closely than the unfurled test angle.

Five-angle check. The C_4 folded distances are 15, 30, 45, 30, and 15 degrees at the five test angles. The observed crossing bands are 15, 30, 30, 30, and 15 degrees: they broadly follow the folded-distance ordering, with the 135° target reached before its folded distance. Full accuracy curves and residual-gap comparisons retain this broad but imperfect agreement (Appendix Table 20). In the initial 150° comparison, C_4 rises from 68.48% at $T = 15$ to 77.04% at $T = 30$, then ranges from 77.37% to 77.73% over $T = 45$–180. A C_4 model trained at $T = 45$ exceeds plain trained at $T = 180$ by 3.93 percentage points (paired exploratory 95% interval $[3.20, 4.67]$). Thus the folded-distance identity is useful context for the curve, not a rule fixing the target crossing.

Approximate C_8. An ideal C_8 action folds 150° to 15° and 135° to zero. Interpolated C_8 instead has the same observed crossing bands as C_4; steerable C_8 reaches all targets at $T = 15$, leaving smaller crossings unresolved. The initial two-seed steerable 3×3 cells (Weiler & Cesa, 2019), reported separately from the full grid, also show higher narrow-band accuracy than interpolation. Yet all recorded post-training 45° errors overlap: 0.376–0.759 for interpolation, 0.176–0.637 for steerable 3×3, and 0.225–0.507 for steerable 5×5 (Table 19). These pooled-feature invariance diagnostics on random inputs do not establish a monotonic error–accuracy relation or a threshold assigning an exact C_8 subgroup.

Earlier small-network CIFAR curves and Tiny-ImageNet/Galaxy10 endpoint checks use separate protocols. Their results remain in Appendix E; they are not pooled with either coverage replication.

4 When do learned-prior gains persist?

The coverage experiments concern constructed group structure. We next compare the learned orientation prior GCLS with augmentation on flat and hierarchical backbones. These task-specific controls concern learned alignment, not methods that explicitly regularize equivariance.

Learned-prior probe (GCLS). GCLS combines orientation-constrained sampling, where a learned local orientation $\theta(x)$ applies a single-parameter rotation-like adjustment to a neighborhood sampling template, with channel-wise low-rank alignment parameterized by an input-dependent antisymmetric generator $M = AB^\top - BA^\top$. These operations do not guarantee input-rotation equivariance; the probe represents this learned alignment mechanism, not all forms of soft equivariance.

4.1 Flat-network controls

Classification (small nets). On a shallow flat CNN ($\sim$158k parameters) under severe elastic out-of-distribution (OOD) shift (three seeds), the prior scores 52.7 versus 47.5 for the no-alignment

experiment	arch	prior−plain	aug−plain	aug−prior
CIFAR ResNet 0.7M	strong	$+0.10$ (p=.66)	$+6.99$ (p=.003)	$+6.89$
CIFAR ResNet 4.4M (6× scale ctrl)	strong	$+0.57$ (p=.49)	$+6.54$ (p=.003)	$+5.97$
CIFAR ViT 0.5M (prior $+157$k params)	strong	$+1.10$ (p=.09)	$+4.58$ (p=.001)	$+3.48$
Tiny-ImageNet ResNet (clean 60.2%)	strong	$+0.52$ (p=.04)	$+11.45$ (p<.001)	$+10.93$
ImageNet-100 ResNet-50 (clean 79.5%)	strong	-0.45 (p=.34)	$+2.28$ (p=.01)	$+2.73$
RAFT flow (OOD EPE, σ units)	strong	$+0.33\sigma$ (n.s.)	lower EPE	rot.-aug. lower
U-Net registration (syn / real)	strong	-0.11σ / $+0.10\sigma$	lower error	augmentation lower
flat CNN (§4)	weak	$+\mathbf{5.2}$	≈ 0	-5.2

Table 4: Augmentation comparisons in the tested backbones. Classification experiments use severe elastic out-of-distribution (OOD) shifts and three seeds; reported paired tests have df=2 and are exploratory. Augmentation exceeds the prior by 2.7 to 10.9 percentage points in the hierarchical classification experiments. Flow and registration use task-specific error metrics and are not included in that percentage-point range.

baseline; elastic augmentation scores 47.2. On medical data under parameter matching, BloodMNIST is 76.4 ± 5.2 versus 57.3 ± 5.7 (plain) and 58.4 ± 1.0 (aug), while TissueMNIST is 59.0 ± 2.1 versus 49.1 ± 0.6 and 50.2 ± 0.3. In these small-network comparisons, the prior also outperforms a plain CNN widened $6.8\times$ and trained with augmentation.

Dense field regression. Predicting per-pixel displacement between image pairs with a $\sim$64k-parameter flat CNN gives OOD endpoint errors (three seeds) of 4.53 ± 0.15 for plain, $\mathbf{3.63 \pm 0.13}$ for the prior, 4.72 ± 0.14 for rotation-aug, and 4.54 ± 0.11 for elastic-aug. The prior improves on plain by 0.90 in endpoint error (EPE) and outperforms both augmentation baselines.

Capacity scan. As the flat CNN grows from 64k to $\sim$3M parameters, the prior's EPE advantage over elastic augmentation persists ($+0.94 \rightarrow +0.96 \rightarrow +1.10$). Parameter count alone therefore does not by itself account for the smaller gains on hierarchical backbones below.

4.2 Hierarchical backbones and augmentation

Table 4 compares plain, +prior, and +augmentation across architectures and scales. On hierarchical classification backbones, the prior gains -0.45 to $+1.10$ percentage points over plain, with several inconclusive three-seed comparisons. In contrast, **augmentation exceeds the prior by 2.7 to 10.9 percentage points**. Flow and registration show the same ordering in their respective metrics. On the flat CNN, the prior gains 5.2 percentage points while augmentation gains approximately zero.

Three controls help interpret this difference. *Data scarcity*: reducing U-Net registration data from 2000 to 50 pairs at fixed gradient steps keeps the prior's margin within $\pm 0.35\sigma$. Together with the capacity scan, this limits explanations based solely on data quantity or parameter count; architecture and task still differ. *Component isolation*: channel alignment alone shows no measured gain over identity. *Distortion control*: rotation augmentation improves RAFT under global rotation.

4.3 Six geometric-prior probes in a common ResNet setup

We insert five established geometric mechanisms into the same parameter-matched ResNet experiment, each as one stage after the stem (Appendix Table 25). This compares local modules at a fixed location; full-network implementations are considered below.

Gains range from 0.06 percentage points for CoordConv to 4.14 percentage points for non-rigid TPS, but every insertion remains below augmentation. The mechanisms differ in several respects, precluding an explanation based on expressivity alone. Complementarity requires a prior-plus-augmentation arm.

4.4 Matched tasks and full-network implementations

STN and full-network ORN both retain gains over augmentation on tasks suited to their geometric structure, providing complementary task-matched context for the local ResNet comparison.

STN on affine distortion. With a shallow CNN and affine rotation-translation-scale (RTS) distortion, STN exceeds a plain CNN trained on the same distortion distribution by 23.3 percentage points (72.2 vs 48.9). This benefit concerns learned alignment within the shared training and test support, rather than extrapolation.

Full-network ORN. ORN (Zhou et al., 2017) with an Active Rotating Filter at every layer (N=8 orientations) and final ORPooling scores 95.0% on rotated-MNIST after training on $\pm 30^\circ$ and testing over $\pm 180^\circ$ (Table 28). It gains 38.4 percentage points over plain and 2.5 percentage points over full-angle augmentation (exploratory paired $p = .002$), close to the C_4 control. ARF rotates kernels and cyclically shifts orientation channels throughout the network. Feature-invariance error is 1.4×10^{-6} at the grid-exact 90° action; interpolated 45° actions remain approximate. Both implementation depth and task differ from the one-stage probe, leaving their individual effects unresolved.

4.5 Routing and matched rotations for the learned orientation prior

Routing. Three conditional-routing designs (per-pixel gated dual-head, sample-level mixture-of-experts (MoE), residual-gated gauge) gain 0.5–1.3 percentage points over plain but remain 5.7–6.2 percentage points below augmentation. Reported routing weights agree to three decimals between clean and severe-OOD inputs, and the deepest residual gate learns $\alpha = -0.94$. These aggregate diagnostics reveal little adaptation to the tested distribution shift, although they cannot resolve behavior on individual subsets.

Matched-rotation test. On rotated-MNIST, the orientation prior addresses the same rotation family used for evaluation. Training on $\pm 30^\circ$ and testing over $\pm 180^\circ$ gives a gauge-minus-plain mean difference of 0.0 percentage points (both 58.3%). Adding gauge to full-angle augmentation changes accuracy by -0.11 percentage points. Matching the transformation family is therefore insufficient to produce an extrapolation gain for this learned prior.

5 Constructed equivariance as a positive control

We evaluate a C_4 group-equivariant CNN (Cohen & Welling, 2016) as a positive control, using grid-exact 90° rotations via `rot90` and orientation pooling. Its pooled feature changes by $\sim 10^{-6}$ under a 90° input rotation (plain CNN: 5.6×10^{-2}), providing a reference with verified right-angle invariance for the learned-prior comparisons.

Extrapolation. Narrow-band arms train on $\pm 30^\circ$ and test over $\pm 180^\circ$, including unseen angles (3 seeds). On rotated-MNIST, the learned-prior experiment reports plain 58.3, gauge 58.3, and full-angle rotaug 96.1. A separate C_4 experiment with different model sizes reaches **95.1**, a 38.1-percentage-point gain over its own plain baseline; the two experiments are not jointly parameter-matched. On rotated-CIFAR, a single experiment reports plain 42.7, gauge 41.7, C_4 **66.0**, and rotaug 55.7. The C_4 gains are 23.3 percentage points over plain and 10.3 percentage points over full-angle augmentation under these budgets, while gauge is 1.0 percentage point lower than plain.

Scale and group controls. At approximately $6\times$ the model size, the gap between the hard control and plain remains similar (22.65 versus 23.3 percentage points), whereas the separate learned-prior architecture controls reduce the prior's gain from 5.2 to 0.6 percentage points. These comparisons change multiple properties and cannot isolate capacity. D_4 (67.9) and interpolated C_8 (67.5) have similar accuracy to C_4 (66.0) on rotated-CIFAR; all share a grid-exact C_4 subgroup. Interpolated C_8 has feature-invariance error 3.5×10^{-6} at 90° but approximately 0.4 at 45°. These measurements motivate the generator-level audit but do not validate a scalar threshold for assigning effective subgroups.

Task controls. On rotated Tiny-ImageNet, arms trained on $\pm 30^\circ$ are tested over $\pm 30^\circ$ and $\pm 180^\circ$. Accuracy changes by 0.2 percentage points for C_4 and by -0.3 percentage points for approximate C_8. Gauge starts highest (47.4) but drops by 4.5 points, close to the plain baseline's 4.3-point drop: higher accuracy does not imply a smaller rotation penalty. Full-angle augmentation also changes little across bands but has lower absolute accuracy here. On galaxy morphology, no arm loses more than approximately 0.6 percentage points, consistent with the dataset's lack of a canonical orientation.

Feature-level check. A transformation-error diagnostic motivated by Gruver et al. (2023) increases during training for both plain (0.20→0.32) and gauge (0.34→0.36), while the constructed control remains near 10^{-6}. Gauge's error is higher than plain's at both checkpoints. Thus its accuracy gains do not coincide with improved invariance in the probed features and transformations.

6 A DIAGNOSTIC DECOMPOSITION OF OUT-OF-DISTRIBUTION (OOD) GAINS

For a prior P and a matched baseline B, both trained in-band and evaluated both in-band and out of distribution, the out-of-distribution accuracy difference has the following additive decomposition:

$$\text{capacity}(P) = \text{acc}_{\text{in}}(P) - \text{acc}_{\text{in}}(B), \qquad \text{robustness}(P) = \text{drop}(B) - \text{drop}(P), \tag{3}$$

where $\text{drop}(X) = \text{acc}_{\text{in}}(X) - \text{acc}_{\text{OOD}}(X)$. The terms sum exactly to $\text{acc}_{\text{OOD}}(P) - \text{acc}_{\text{OOD}}(B)$. Here "capacity" labels the in-distribution accuracy gain and "robustness" the reduction in accuracy drop; neither label identifies a causal mechanism. In particular, a smaller drop alone establishes neither equivariance nor an advantage over augmentation. Appendix Table 27 and Fig. 11 report the values.

On the Tiny-ImageNet rotation testbed, gauge has an in-distribution gain of 5.8 percentage points and a robustness term of -0.2 percentage points (the drop increases by 0.2 points): its higher OOD accuracy is associated with a higher starting accuracy, not with a smaller measured rotation penalty. The C_4 and approximate C_8 controls reduce that penalty by 2.8 and 3.3 percentage points; the rotated-CIFAR controls yield 24 to 25 percentage-point reductions with in-distribution differences near zero. These contrasts motivate measuring both endpoints instead of interpreting absolute OOD accuracy alone.

The elastic-OOD results are less sharply separated: soft priors reduce the drop by 0.6 to 1.0 percentage points, and STN reaches 2.1 percentage points. Assessing whether these gains survive augmentation requires a stacked arm on the same task. The -0.11 percentage-point gauge-plus-augmentation result in §4.5 concerns rotated-MNIST and does not answer that question for elastic deformation. The decomposition identifies which endpoint changes; explicit scaling and augmentation controls establish whether the gain survives those alternatives.

7 DISCUSSION AND LIMITATIONS

Coverage savings depend on the budget, target, and dataset. Exact C_4 reaches the fixed-epoch plain reference with less angular support on both ImageNet-100 and CIFAR-10. With matched training time and four-view averaging, the ImageNet-100 baseline recovers that reference at 10 of 11 angles using $T = 30$, while C_4 retains 3–4 percentage points. On CIFAR-10, substantial narrow-band gains shrink to near-zero broad-band differences and reverse under zero-filled rotation. Reporting the full curves and censored crossings makes this boundary visible.

Both coverage tasks assume rotation-label preservation and use low-resolution images, three seeds, sparse bands, and one schedule per budget. The alternative rotation changes sampling and boundaries; ImageNet-100 also changes evaluation precision. Matched-time pairs have separate measured budgets at each band, and inference costs differ. The complete CIFAR-10 sweep does not resolve the missing intermediate ImageNet-100 controls or demonstrate transfer to natural domain shifts. Target crossings near a plateau are sensitive to seed variation. The folded-distance pattern is empirical and imperfect; all tested C_8 actions remain approximate.

Separate learned-prior controls show larger augmentation gains on hierarchical backbones and useful prior gains on flat networks and matched tasks. Data-scarcity and widening controls do not reproduce this pattern, leaving architecture and task as unresolved factors. The evidence does not show that data scale generally removes soft-equivariance benefits.

REFERENCES

Ahmed M. Adly. Measuring the symmetry–data exchange rate. *arXiv preprint arXiv:2606.01090*, 2026.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv preprint arXiv:2410.23179*, 2024.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision*, 2017.

Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450 (2):1441–1459, 2015.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, 2021.

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. In *International Conference on Learning Representations*, 2023.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, 2018.

Sneh Pandya, Purvik Patel, Ondrej Franc, and Jonathan Blazek. E(2) equivariant neural networks for robust galaxy morphology classification. *arXiv preprint arXiv:2311.01500*, 2023.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems*, 2022.

Anna M. M. Scaife and Fiona Porter. Fanaroff-Riley classification of radio galaxies using group-equivariant convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 503(2):2369–2379, 2021.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, 2022.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, 2019.

Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

A USE OF LARGE LANGUAGE MODELS

We used language models to aid or polish the writing. The authors wrote and verified the core experiments and take responsibility for the paper.

B FULL NUMBERS AND REPRODUCIBILITY

Results are backed by per-experiment JSON files and logs. Experiment scripts record comparison criteria and per-seed results; script headers alone do not constitute external preregistration.

Coverage provenance. The completed ImageNet-100 follow-up contains 60 unique $(\text{model}, T, \text{seed})$ runs: four arms, five coverage levels, and three seeds, with five angle evaluations in every run. Duplicate lane copies were checked by SHA-256 and counted once; no conflicting values were found. Table 17 gives the earlier headline cells, while Table 20 gives the complete five-angle coverage-saving analysis. Paired contrasts use only common seed IDs. These small-n intervals assume approximately normal, independent training-seed effects; they are not corrected for multiple comparisons and exclude model-selection and dataset uncertainty. The derived CSV and JSON audit retain source filenames, hashes, duplicate records, and full-precision observations.

Generator diagnostic. Coverage scripts compute $\|F(x) - F(R_\alpha x)\|_F / (\|F(x)\|_F + 10^{-9})$ on a batch of four Gaussian random inputs, where F is the spatially and orientation-pooled feature in evaluation mode after training. This measures feature invariance, not layerwise equivariance or a worst-case bound. Grid-aligned angles use `rot90`; $45°$ uses bilinear sampling with border padding and `align_corners=False`. Initialization values mentioned in script headers are not imputed into the post-training audit. Error ranges in Table 19 include every stored value, rather than only values near 0.3 or 0.5.

Data and budget. The cache builder resizes decoded RGB images to 128×128, then randomly partitions each class with a fixed NumPy seed of 0 into 100 held-out images and the remainder for training. Coverage runs average-pool to 64px and use ImageNet channel normalization. Each training batch receives rotations sampled uniformly from $[-T, T]$; evaluation uses bilinear rotation with border padding, including at zero angle and at $105/120/135/150/165°$. Models train from scratch with automatic mixed precision (AMP), SGD (learning rate 0.2, momentum 0.9, Nesterov, weight decay 5×10^{-4}), OneCycle scheduling (`pct_start=0.1`), label smoothing 0.1, batch 512, and the final checkpoint after 30 epochs. The held-out split is fixed; rotation label preservation is an assumption of this synthetic evaluation. Parameter matching is approximate; Table 21 adds single-step FLOP and peak-memory references.

Architectures. Plain ResNet-18 uses a 3×3, stride-1 stem without the initial max-pool. The C_4 model has two residual blocks per stage at widths 32/64/128/256, stride-1 group convolutions, external 2×2 spatial max-pooling between stages, and orientation max-pooling before global spatial averaging. These architectural differences accompany parameter matching and should be retained when reproducing the comparison.

Stacked augmentation follow-up. The companion runner uses the same 64px cache, split, optimizer, batch size, and 30-epoch budget at $T \in \{15, 30, 180\}$. Both arms apply independently sampled integer shifts in $[-4, 4]$ along each axis using reflection padding, followed by a per-sample horizontal flip with probability 0.5, after rotation. Evaluation applies rotation only. Three seeds are evaluated at the original five angles. The $T = 15$ paired interaction is reported in Table 5; the expanded grid and post-evaluation appear in Appendix C. These runs remain separate from the frozen 60-run coverage grid.

Convergence audit. An independent 60-epoch continuation of the exact C_4, $T = 30$, seed-0 run reaches 78.53% at $150°$, versus 77.44% at the fixed 30-epoch checkpoint. This $+1.09$ point diagnostic is not mixed into the 30-epoch coverage statistics; it records that the fixed-budget comparison is reproducible but not fully converged.

Test angle	Plain+aug	C_4+aug	Δplain	ΔC_4	Interaction
105°	41.09	79.97	+3.02	+2.42	-0.60
120°	36.42	70.48	+2.43	+0.85	-1.58
135°	34.13	63.91	+1.46	+0.79	-0.67
150°	36.78	70.90	+1.60	+2.23	+0.63
165°	43.17	80.19	+2.32	+2.86	+0.54

Table 5: Independent stacked-arm follow-up at $T = 15$ on ImageNet-100 (three seeds). Plain+aug and C_4+aug use the same train-only reflected four-pixel translation and 0.5 horizontal-flip augmentation; evaluation remains rotation-only. Δ columns are paired gains over the corresponding frozen-grid arm, and Interaction is $(\Delta C_4 - \Delta\text{plain})$. Values are percentage points; this T=15 companion result is excluded from the 60-run coverage-saving estimand.

Scripts by section:

- `train_pathmnist_ablations.py`, `aug_control_*` — small-net classification (§4)
- `keystone_flow.py`, `keystone_capacity_scan.py` — field regression and capacity scan (§4)
- `resnet_gate.py`, `vit_gate.py` — classification comparisons (§4.2)
- `raft_injection_gate.py`, `unet_reg_gate.py` — dense-task comparisons (§4.2)
- `real_unet_reg_gate.py`, `unet_scarcity_scan.py` — registration controls (§4.2)
- `ortho_align_gate.py` — component isolation (§4.2)
- `bigresnet_gate.py` — backbone scale control (§4.2)
- `in100_build_cache.py`, `in100_absorption_gate.py` — ImageNet-100 control (§4.2)
- `lit_priors_gate.py` — five adopted soft priors (§4.3)
- `gateA_stn_reeval.py`, `orn_audit_gate.py` — home-turf controls (§4.4)
- `resnet_dualhead_gate.py`, `moe_gate.py`, `residual_gauge.py` — routing (§4.5)
- `groupmatch_rotmnist_gate.py` — matched-rotation control (§4.5)
- `gcnn_equiv_gate.py`, `rotcifar_equiv_gate.py` — C_4 controls (§5)
- `bigrotcifar_gate.py`, `d4_equiv_gate.py`, `c8_equiv_gate.py` — scale and group controls (§5)
- `lie_equiv_probe.py` — feature-level check (§5)
- `orbit_coverage_sweep.py` — CIFAR coverage sweep (§3)
- `in100_coverage_gate.py` — ImageNet-100 coverage and wider-plain controls (§3.1)
- `c8_coverage_gate.py`, `c8_steerable_gate.py` — approximate C_8 variants (§3.1)
- `in100_coverage_merge.py` — coverage merger; script-header predictions are internal records, not external preregistrations
- `run_aug_followup.py`, `in100_matched_time.py` — augmented coverage and paired training-time controls
- `run_aug_posteval.py`, `analyze_aug_posteval.py` — additional angles, test-time averaging, rotation sensitivity, and batched inference timing

C POST-HOC CHECKPOINT EVALUATION

After all training runs were complete, we evaluated the 24 stored checkpoints on the fixed 10,000-image ImageNet-100 test split. This exploratory follow-up was specified after observing the training results and was not externally preregistered. It adds six fixed angles, $97.5/112.5/127.5/142.5/157.5/172.5°$, to the original five, evaluates plain models with four-view right-angle probability averaging (TTA_4), and evaluates plain checkpoints trained for the measured

	original five angles			new six angles		
T	plain	plain+TTA$_4$	exact C_4	plain	plain+TTA$_4$	exact C_4
fixed 30-epoch training						
15	38.32 ± 0.18	62.27 ± 0.24	73.09 ± 0.20	38.91 ± 0.24	63.85 ± 0.04	74.61 ± 0.09
30	43.76 ± 0.17	70.47 ± 0.27	79.19 ± 0.05	44.05 ± 0.13	70.98 ± 0.24	79.42 ± 0.07
180	73.65 ± 0.14	74.84 ± 0.17	79.33 ± 0.17	73.62 ± 0.23	74.92 ± 0.18	79.33 ± 0.24
plain trained for the matched C_4 wall-clock budget						
30	48.64 ± 0.12	75.16 ± 0.11	79.19 ± 0.05	48.95 ± 0.16	75.42 ± 0.08	79.42 ± 0.07
180	75.51 ± 0.35	76.36 ± 0.19	79.33 ± 0.17	75.57 ± 0.24	76.30 ± 0.13	79.33 ± 0.24

Table 6: Post-hoc checkpoint evaluation on ImageNet-100 (accuracy in percent, mean $\pm$ sample SD over three seeds). Each seed first averages the indicated angles. The upper block compares fixed 30-epoch checkpoints; the lower block makes the direct same-T comparison after training plain models for the hardware-specific wall-clock budget of their paired C_4 run. TTA$_4$ averages probabilities over four right-angle views. All entries use the original border-padded rotation operator.

wall-clock budget of the paired exact-C_4 run. Every reported cell contains all three seeds. These are further evaluations of one task and split, so they test angular and implementation sensitivity but do not demonstrate generalization to a new task.

The primary operator reproduces the original evaluation implementation: bilinear rotation with border padding and its grid construction under autocast. Training uses the same bilinear/border operation outside autocast. The alternative uses torchvision bilinear rotation in float32 with zero fill in normalized space. It therefore changes sampling implementation, padding, and numerical precision together; it is a sensitivity check and does not isolate any one factor or constitute a natural domain shift. TTA$_4$ averages float32 softmax probabilities after applying exact right-angle rotations to the already test-rotated input. The timing benchmark uses the same local RTX 4080, batch 512, AMP, three warmups, and ten synchronized repetitions.

Matched-time budgets exclude setup, evaluation, and checkpoint I/O in both arms. The C_4 epoch-level timer includes scheduler work, whereas plain uses synchronized batch-level timers that exclude scheduler bookkeeping and may exceed the budget by at most one batch. These are measured training-time comparisons rather than identical end-to-end runtimes. The added training and evaluation runs use PyTorch 2.7.0+cu128 and torchvision 0.22.0+cu128; hardware pairings and fractional epoch counts appear in Table 10.

D FULL CIFAR-10 REPLICATION WITH PAIRED TRAINING BUDGETS

This follow-up applies the coverage protocol to all 50,000 training and 10,000 test images in CIFAR-10 at their native 32×32 resolution. It is separate from the small-network Rot-CIFAR10 experiments elsewhere in this appendix. The same training examples are used at each sampled half-width $T \in \{15, 30, 45, 90, 180\}^\circ$; only the distribution of sampled training rotations changes. There are three seeds and three training arms at every T, giving 45 completed runs. The protocol, source files, and dataset identity were frozen before this replication began, after the ImageNet-100 results were available. This is a prospective follow-up, not an external preregistration. It tests transfer to a second image dataset under synthetic rotations; it does not establish generalization to another transformation family or a natural domain shift.

Architectures and training. The plain arm uses a ResNet-18 with a 3×3, stride-one stem and no initial max-pooling. The exactly implemented C_4 arm uses the same residual architecture family, widths $32/64/128/256$, external 2×2 pooling, and orientation max-pooling. Their parameter counts are 11.174M and 11.168M, respectively. Fixed-epoch runs train for 30 epochs with batch size 512, AMP, SGD (maximum learning rate 0.2, momentum 0.9, Nesterov, weight decay 5×10^{-4}), label smoothing 0.1, and a OneCycle schedule with a 10% warm-up fraction. All arms use reflection-padded integer translations in $[-4, 4]$ pixels along each axis and horizontal flips with probability 0.5. Image normalization uses training-split statistics only. Training rotations are uniform on $[-T, T]$, using bilinear sampling and border padding. The final budget checkpoint is evaluated without checkpoint selection on test results.

operator	angle	target	plain	plain+TTA$_4$	exact C_4	time-matched TTA$_4$ at $T = 30$
border	105°	73.74	180	180	15	reaches
	120°	73.59	180	180	30	reaches
	135°	73.33	180	180	30	misses
	150°	73.56	180	180	30	reaches
	165°	74.01	180	180	15	reaches
	97.5°	73.78	180	180	15	reaches
	112.5°	73.76	180	180	15	reaches
	127.5°	73.24	180	180	30	reaches
	142.5°	73.30	180	180	30	reaches
	157.5°	73.66	180	180	15	reaches
	172.5°	73.98	180	180	15	reaches
zero-fill	105°	72.90	180	180	15	reaches
	120°	72.02	180	180	30	reaches
	135°	71.83	180	180	30	misses
	150°	71.86	180	180	30	reaches
	165°	72.83	180	180	15	reaches
	97.5°	73.06	180	180	15	reaches
	112.5°	72.36	180	180	15	reaches
	127.5°	71.69	180	180	30	reaches
	142.5°	71.75	180	180	30	reaches
	157.5°	72.37	180	180	15	reaches
	172.5°	73.33	180	180	15	reaches

Table 7: Shared-target crossings on the explicit sparse grid $T \in \{15, 30, 180\}$. For each operator and angle, the target is the same-operator mean accuracy of fixed-epoch plain training at $T = 180$; columns give the first sampled fixed-epoch T reaching it, without interpolation. The time-matched plain+TTA$_4$ checkpoint at $T = 30$ reaches this angle-specific target at 10 of 11 angles under each operator, missing only 135° (by 0.88 points with border padding and 1.33 points with zero fill). This is an observed $T = 30$ result, not a crossing estimate: matched-time $T = 15$ was not evaluated. Consequently, the fixed-epoch $156^\circ/160^\circ$ mean sampled savings of C_4 do not survive unchanged under the strongest compute-and-inference control.

operator	cohort	$T = 30$	$T = 180$
border	original	$+4.03 \pm 0.16$	$+2.97 \pm 0.33$
border	new	$+4.00 \pm 0.13$	$+3.03 \pm 0.37$
zero-fill	original	$+4.00 \pm 0.15$	$+3.02 \pm 0.26$
zero-fill	new	$+3.99 \pm 0.06$	$+2.96 \pm 0.12$

Table 8: Paired exact-C_4 minus time-matched plain+TTA$_4$ accuracy in percentage points. Pairing uses seed labels, although the arms have independent optimization randomness. Each seed is averaged over angles before the mean and sample SD across three seeds.

Training-time control and inference augmentation. The third arm trains plain for the CUDA-synchronized training seconds consumed by the C_4 run at the same T and seed on the same GPU and software stack. Its OneCycle schedule is indexed by elapsed training time. Both arms time the same per-batch region: permutation/indexing, normalization, augmentation, forward/backward computation, and optimizer/scaler updates. Setup, checkpoint writing, evaluation, and scheduler bookkeeping are excluded in both arms. This gives a pairwise training-time control at each T, not a single globally constant time budget over the coverage sweep. The final batch may overshoot the paired budget by at most one batch. Plain is evaluated both with one view and with probability averaging over four exact right-angle rotations (TTA$_4$); C_4 is evaluated with one view. Training-time matching does not match inference cost. Tables 15 and 16 report both costs.

Evaluation and audit. The primary cohort is $\{105, 120, 135, 150, 165\}^\circ$ and the separate additional cohort is $\{97.5, 112.5, 127.5, 142.5, 157.5, 172.5\}^\circ$. An unrotated evaluation is retained as a diagnostic and excluded from coverage savings. Each test angle is evaluated with the training bilinear border operator and, separately, with torchvision bilinear rotation using zero fill in normalized space. These operators are not pooled. Pooled-feature invariance to right-angle rotations is audited before and after training in FP32 with TF32 disabled, on fixed natural-image and Gaussian inputs, using relative ℓ_2 error tolerance 10^{-4}. Source/configuration identities, all evaluation counts, and the complete coverage–seed–arm grid are checked before aggregation.

inference method	ms/image
plain	0.087
plain+TTA$_4$	0.348
exact C_4	0.470

Table 9: Median synchronized inference time per image on one RTX 4080 at batch 512. Timings include right-angle transforms and probability aggregation for TTA$_4$ and exclude the input test rotation. These are amortized batch measurements, not single-image request latencies.

device (seeds)	T	paired budget (s)	plain epochs
RTX 4080 (0)	30	5500.813	172.338
RTX 4080 (0)	180	5503.010	172.548
RTX PRO 6000 (1,2)	30	2774.329–2774.758	216.825–217.004
RTX PRO 6000 (1,2)	180	2775.350–2776.068	217.000–217.022

Table 10: Matched-time training provenance. Each plain run uses the CUDA-synchronized training-time budget of the same-seed, same-T exact-C_4 run on the same host and GPU. OneCycle progress is indexed by elapsed budget, so the resulting epoch counts differ by hardware. Accuracy contrasts combine hardware across seeds; runtime speedups are not pooled.

Statistics and crossing definitions. For each macro accuracy, angles are first averaged within a training seed, then mean and sample SD are computed over the three independent training seeds. Paired gaps match seed labels; the labels do not imply identical optimization randomness between architectures. Test angles are repeated measurements of each trained model, not independent replicates. Each crossing target is the three-seed mean accuracy at $T = 180^\circ$ of either the fixed-epoch single-view plain reference or the matched-time plain+TTA$_4$ reference at that same test angle and operator. The reference crossing is measured on the grid, rather than assumed to equal 180°. Both architectures use the first sampled crossing, without interpolation or monotonic fitting. Missing crossings are right-censored beyond the sampled grid and have no numerical saving. A crossing of cohort-average accuracy is a separate quantity from the average of per-angle crossings.

What the replication supports. At fixed epochs, the mean per-angle half-width savings are 156° for the original five angles and 160° for the additional six under both operators. The stronger reference changes that conclusion. Under border padding, the original-five mean per-angle saving is 42°, but the crossing of their macro accuracy gives zero saving: both mean curves first attain the matched-time reference macro target at $T = 180^\circ$. The additional-six macro also gives zero saving, while one of its individual C_4 crossings is censored. Under zero fill, three of five original and four of six additional C_4 crossings are censored, as are both macro crossings. Thus the complete grid supports a coverage-dependent advantage at narrow training support, together with sensitivity to the accuracy target, budget, aggregation order, and image operator; it does not support universal coverage savings against the strongest plain baseline. The 42° figure is a descriptive mean-curve statistic, not a three-seed confidence estimate.

		30 epochs			matched-time plain		
angles	T	plain	plain+TTA$_4$	C_4	single	TTA$_4$	paired gap
original five	15	33.65 ± 0.33	59.33 ± 0.62	81.99 ± 0.64	35.90 ± 0.49	62.29 ± 0.71	$+19.70 \pm 1.29$
	30	38.67 ± 0.46	71.79 ± 0.19	90.70 ± 0.10	41.55 ± 0.50	75.86 ± 0.49	$+14.83 \pm 0.56$
	45	42.25 ± 0.42	79.82 ± 0.15	91.90 ± 0.13	46.07 ± 0.19	84.05 ± 0.34	$+7.85 \pm 0.28$
	90	56.01 ± 0.42	86.85 ± 0.04	92.00 ± 0.06	65.85 ± 0.52	91.94 ± 0.06	$+0.07 \pm 0.09$
	180	84.66 ± 0.23	86.11 ± 0.20	92.13 ± 0.09	91.09 ± 0.18	92.02 ± 0.11	$+0.11 \pm 0.02$
additional six	15	34.65 ± 0.33	60.36 ± 0.54	83.87 ± 0.59	36.85 ± 0.30	63.03 ± 0.70	$+20.84 \pm 1.24$
	30	38.75 ± 0.41	71.18 ± 0.23	91.21 ± 0.08	41.91 ± 0.42	74.96 ± 0.46	$+16.25 \pm 0.42$
	45	42.25 ± 0.51	78.33 ± 0.15	91.93 ± 0.14	46.60 ± 0.32	82.57 ± 0.50	$+9.35 \pm 0.38$
	90	57.17 ± 0.43	86.87 ± 0.07	91.97 ± 0.08	66.85 ± 0.49	91.99 ± 0.01	-0.02 ± 0.09
	180	84.54 ± 0.29	86.03 ± 0.31	92.04 ± 0.08	91.07 ± 0.11	92.02 ± 0.11	$+0.02 \pm 0.03$

Table 11: CIFAR-10 cohort-average accuracy under border padding. Each seed first averages the stated angles; entries show mean $\pm$ sample SD over three seeds. The paired gap is single-view C_4 minus matched-time plain+TTA$_4$. The fixed-epoch plain+TTA$_4$ and matched-time single-view columns separate inference augmentation from additional training.

		30 epochs			matched-time plain		
angles	T	plain	plain+TTA$_4$	C_4	single	TTA$_4$	paired gap
original five	15	32.78 ± 0.48	57.88 ± 1.32	76.43 ± 0.67	35.85 ± 0.20	61.20 ± 0.40	$+15.24 \pm 0.69$
	30	37.89 ± 0.53	69.03 ± 0.41	87.94 ± 0.21	41.49 ± 0.49	74.31 ± 0.47	$+13.62 \pm 0.63$
	45	41.15 ± 0.18	76.38 ± 0.34	89.60 ± 0.21	45.49 ± 0.14	81.88 ± 0.38	$+7.72 \pm 0.18$
	90	54.30 ± 0.40	83.08 ± 0.17	89.72 ± 0.09	64.66 ± 0.67	89.94 ± 0.19	-0.22 ± 0.25
	180	80.25 ± 0.42	81.79 ± 0.43	89.91 ± 0.05	89.01 ± 0.20	90.28 ± 0.09	-0.37 ± 0.11
additional six	15	33.86 ± 0.46	59.27 ± 1.20	79.48 ± 0.64	36.75 ± 0.25	62.12 ± 0.52	$+17.36 \pm 0.72$
	30	38.03 ± 0.48	68.50 ± 0.31	88.76 ± 0.10	41.75 ± 0.46	73.61 ± 0.45	$+15.15 \pm 0.55$
	45	41.25 ± 0.26	75.26 ± 0.32	89.89 ± 0.15	45.76 ± 0.02	80.64 ± 0.41	$+9.25 \pm 0.34$
	90	55.53 ± 0.25	83.50 ± 0.13	90.01 ± 0.12	65.83 ± 0.61	90.41 ± 0.19	-0.39 ± 0.29
	180	80.73 ± 0.44	82.33 ± 0.46	90.17 ± 0.08	89.24 ± 0.14	90.47 ± 0.13	-0.30 ± 0.06

Table 12: CIFAR-10 cohort-average accuracy under torchvision zero fill. Each seed first averages the stated angles; entries show mean $\pm$ sample SD over three seeds. The paired gap is single-view C_4 minus matched-time plain+TTA$_4$. The fixed-epoch plain+TTA$_4$ and matched-time single-view columns separate inference augmentation from additional training.

	fixed-epoch target				matched-time+TTA$_4$ target			
test angle	target	plain $T^\star$	C_4 $T^\star$	saving	target	plain $T^\star$	C_4 $T^\star$	saving
Original five angles								
105°	84.80	180	15	165	92.29	90	30	60
120°	84.47	180	30	150	91.79	180	90	90
135°	84.50	180	30	150	91.93	180	180	0
150°	84.40	180	30	150	91.92	180	180	0
165°	85.14	180	15	165	92.15	90	30	60
cohort macro	84.66	180	30	150	92.02	180	180	0
Additional six angles								
97.5°	84.55	180	15	165	92.14	90	15	75
112.5°	84.45	180	15	165	92.03	180	30	150
127.5°	84.40	180	30	150	91.88	180	180	0
142.5°	84.26	180	30	150	91.98	180	180	0
157.5°	84.73	180	15	165	92.03	180	> 180	–
172.5°	84.87	180	15	165	92.09	90	15	75
cohort macro	84.54	180	30	150	92.02	180	180	0

Table 13: All mean-curve CIFAR-10 crossings under border padding. Targets are percentages; $T^\star$ and savings are training half-widths in degrees. Each target is the same-angle (or cohort-macro) three-seed mean at $T = 180°$ of its stated plain reference. Plain and C_4 crossings are measured independently on the sampled grid; no interpolation or monotonic fit is used. > 180 marks no observed crossing, with saving unavailable (–) rather than zero. The cohort-macro crossing is calculated after averaging accuracy over angles and differs from the average of per-angle savings.

	fixed-epoch target				matched-time+TTA$_4$ target			
test angle	target	plain $T^\star$	C_4 $T^\star$	saving	target	plain $T^\star$	C_4 $T^\star$	saving
Original five angles								
105°	81.76	180	15	165	90.90	90	30	60
120°	79.42	180	30	150	89.84	180	> 180	–
135°	78.47	180	30	150	89.73	180	> 180	–
150°	79.56	180	30	150	89.87	180	> 180	–
165°	82.03	180	15	165	91.06	90	30	60
cohort macro	80.25	180	30	150	90.28	180	> 180	–
Additional six angles								
97.5°	82.76	180	15	165	91.15	90	15	75
112.5°	80.73	180	15	165	90.38	90	> 180	–
127.5°	78.39	180	30	150	89.70	180	> 180	–
142.5°	78.53	180	30	150	89.73	180	> 180	–
157.5°	81.13	180	15	165	90.43	180	> 180	–
172.5°	82.81	180	15	165	91.45	90	15	75
cohort macro	80.73	180	30	150	90.47	180	> 180	–

Table 14: All mean-curve CIFAR-10 crossings under torchvision zero fill. Targets are percentages; $T^\star$ and savings are training half-widths in degrees. Each target is the same-angle (or cohort-macro) three-seed mean at $T = 180°$ of its stated plain reference. Plain and C_4 crossings are measured independently on the sampled grid; no interpolation or monotonic fit is used. > 180 marks no observed crossing, with saving unavailable (–) rather than zero. The cohort-macro crossing is calculated after averaging accuracy over angles and differs from the average of per-angle savings.

training arm	parameters (M)	FLOPs (G)	single (ms/image)	TTA_4 (ms/image)
plain, 30 epochs	11.174	1.111	0.019	0.073–0.074
C_4, 30 epochs	11.168	4.437	0.110–0.111	–
plain, matched time	11.174	1.111	0.019	0.074

Table 15: CIFAR-10 parameter count, single-view forward FLOPs, and observed inference-time ranges over the 15 jobs per arm on one RTX 4080. FLOPs cover PyTorch-profiler-supported operations and count a multiply-add as two; TTA_4 requires four forwards. Inference uses AMP, batch size 512, three warmups and ten synchronized repetitions; milliseconds per image describe batch throughput, not single-image latency. Input test rotation is excluded; TTA timing includes right-angle rotations and probability averaging.

T	training arm	training seconds	epochs	peak allocated GiB
15	plain, 30 epochs	96.6 ± 0.2	30.0	1.44–1.61
	C_4, 30 epochs	564.4 ± 0.5	30.0	3.09–3.43
	plain, matched time	564.4 ± 0.5	175.1–175.9	1.53–1.70
30	plain, 30 epochs	97.1 ± 0.5	30.0	1.44–1.70
	C_4, 30 epochs	565.1 ± 0.8	30.0	3.25–3.51
	plain, matched time	565.1 ± 0.8	174.4–174.9	1.53–1.78
45	plain, 30 epochs	97.4 ± 0.6	30.0	1.44–1.70
	C_4, 30 epochs	566.8 ± 0.8	30.0	3.26–3.51
	plain, matched time	566.8 ± 0.7	172.4–176.6	1.53–1.78
90	plain, 30 epochs	97.2 ± 0.8	30.0	1.44–1.70
	C_4, 30 epochs	565.2 ± 1.3	30.0	3.26–3.51
	plain, matched time	565.2 ± 1.3	173.7–175.5	1.53–1.78
180	plain, 30 epochs	97.9 ± 0.1	30.0	1.44–1.70
	C_4, 30 epochs	566.4 ± 0.7	30.0	3.26–3.51
	plain, matched time	566.4 ± 0.7	173.8–174.0	1.53–1.78

Table 16: CIFAR-10 training cost by coverage. Training seconds are mean $\pm$ sample SD over three seeds; epochs and memory are observed ranges. CUDA-synchronized timing includes data indexing, normalization, augmentation, forward/backward computation and optimizer updates, while excluding setup, checkpoint I/O, evaluation and scheduler bookkeeping in both arms. Matching is within each seed and coverage, so this is not a globally constant time budget across T. Peak CUDA-allocated memory includes resident train/test tensors, model and optimizer, and excludes allocator-reserved but unused memory.

E Earlier small-network and endpoint checks

In the earlier small-network CIFAR sweep (Fig. 2A), C_4 rises from 54.7% at $T = 15$ to about 65% at larger coverage, while plain rises from 30.8% to 52.3%. Gauge rises from 35.6% to 47.4%; its difference from plain changes from $+4.8$ to -4.9 points. These historical curves use a different protocol from the full CIFAR-10 ResNet-18 experiment in §3.3.

Rotated Tiny-ImageNet and Galaxy10 provide separate three-seed, 22-epoch endpoint checks. At a $\pm 180^\circ$ test band, C_4 reaches 39.36 ± 0.47 versus 34.35 ± 0.64 for plain on Tiny-ImageNet, and 65.89 ± 0.33 versus 59.84 ± 1.06 on Galaxy10. Approximate C_8 reaches 38.16 ± 0.07 and 62.49 ± 0.33, respectively. These checks use different data construction and vary test support; they are not additional training-coverage sweeps.

F Quotient identity and limits of the hypothesis

For an H-invariant classifier f and any $g = hg_T$ with $h \in H$ and $g_T \in G_T$, $f(gx) = f(hg_Tx) = f(g_Tx)$. Thus predictions on the training orbit extend to its H-saturation, HG_T. For circular rotations and a symmetric band, $g \in HG_T$ precisely when $d_H(g) \le T$. This is a standard consequence of invariance, not a new theorem about generalization. It equates predictions of a fixed classifier;

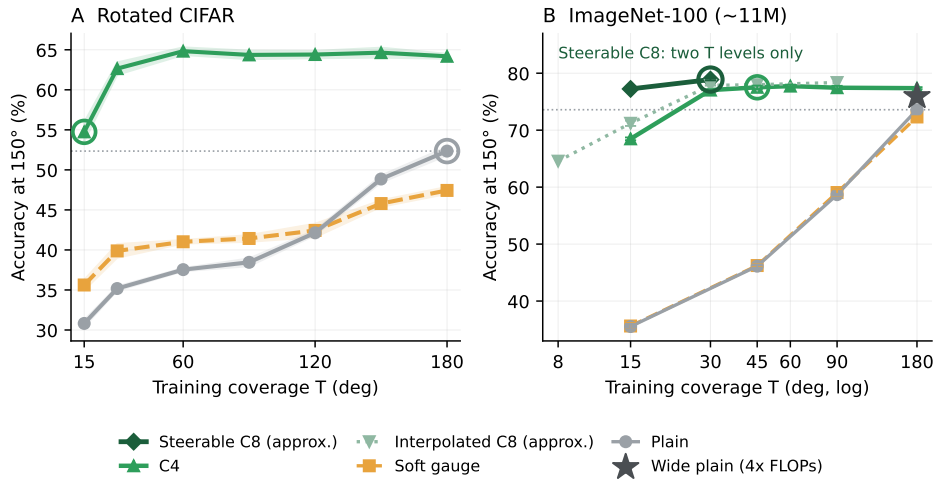


Figure 2: Earlier accuracy curves at a fixed 150° test angle. (A) Small-network rotated CIFAR; (B) the initial ImageNet-100 comparison. Error bars are sample SD. These separate cohorts provide historical context and are not pooled with the full CIFAR-10 replication. The C_8 actions are approximate; wide plain uses about four times plain FLOPs.

it does not equate classifiers trained at different T, guarantee correctness, or specify the plateau of a finite-data accuracy curve. Approximate invariance without a uniform error bound and a classification margin cannot provide the corresponding exact guarantee. The empirical question in this paper is whether folded distance helps organize the measured coverage curves under these practical limitations.

G Coverage results and implementation audits

H Additional tables and figures

Geometric prior evaluation sequence

1. Extrapolation to unseen transformations is required: include a **constructed control**, audit its claimed generator, and evaluate residual coverage.
2. Evaluate a hierarchical backbone and **augmentation**. To assess complementarity, compare prior plus augmentation against augmentation alone.
3. Under architectural or resource constraints, compare the learned prior with the feasible alternatives at the same budget (§4).

Interpretation: Augmentation beating the prior alone establishes a baseline preference in that setting, not redundancy of their combination.

Figure 3: Evaluation sequence motivated by the tested controls. The sequence requires measuring each comparison; it does not predict a new prior's benefit before evaluation.

Arm	T ($\pm^\circ$)	n	Mean	Sample SD	Exploratory 95% CI
Interpolated C_8	8	1	64.56	—	—
Interpolated C_8	15	2	71.21	0.47	[67.02, 75.40]
Interpolated C_8	30	2	77.86	0.01	[77.73, 77.99]
Interpolated C_8	45	1	77.98	—	—
Interpolated C_8	90	1	78.36	—	—
Steerable C_8, 3×3	15	2	77.24	0.28	[74.70, 79.78]
Steerable C_8, 3×3	30	2	78.88	0.15	[77.55, 80.22]
Steerable C_8, 5×5	8	1	69.86	—	—
Steerable C_8, 5×5	15	3	74.06	0.27	[73.40, 74.72]
Steerable C_8, 5×5	30	2	75.97	0.19	[74.26, 77.69]
Steerable C_8, 5×5	45	1	76.61	—	—
Gauge	15	3	35.62	0.55	[34.26, 36.98]
Gauge	45	3	46.23	0.26	[45.58, 46.88]
Gauge	90	3	59.05	0.58	[57.62, 60.48]
Gauge	180	3	72.32	0.13	[72.01, 72.63]
C_4	15	3	68.48	0.25	[67.87, 69.09]
C_4	30	3	77.04	0.15	[76.66, 77.41]
C_4	45	3	77.52	0.30	[76.77, 78.28]
C_4	60	3	77.73	0.26	[77.09, 78.38]
C_4	90	3	77.45	0.33	[76.63, 78.27]
C_4	180	3	77.37	0.18	[76.92, 77.82]
Plain	15	3	35.53	0.48	[34.35, 36.71]
Plain	45	3	46.13	0.28	[45.43, 46.83]
Plain	90	3	58.68	0.14	[58.32, 59.04]
Plain	180	3	73.59	0.14	[73.24, 73.94]
Wide plain	180	3	75.96	0.18	[75.51, 76.40]

Table 17: Initial ImageNet-100 coverage results at test angle 150°, reported separately from the complete 60-run grid. Accuracy is in percent; SD is the sample standard deviation ($n - 1$ denominator). Intervals are exploratory 95% Student-t intervals over seeds, with no multiplicity correction; they are highly uncertain at small n and exclude data-selection uncertainty. SD and intervals are undefined for $n = 1$. Identical copied records count once; conflicting records are excluded and recorded in the audit.

Paired contrast	n	Mean difference	Sample SD	Exploratory 95% CI
C_4@45 − Plain@180	3	3.93	0.29	[3.20, 4.67]
Steerable C_8, 3×3@30 − Plain@180	2	5.32	0.33	[2.34, 8.31]
Steerable C_8, 3×3@15 − Interpolated C_8@15	2	6.03	0.75	[-0.70, 12.76]

Table 18: Paired accuracy differences are reported in percentage points and use only common seed IDs. Intervals are exploratory 95% Student-t intervals over paired seed differences, without multiplicity correction. They do not establish a universal effect.

Arm	T ($\pm^\circ$)	Records	Minimum	Maximum
Interpolated C_8	All	7	0.375524	0.759083
Interpolated C_8	8	1	0.375524	0.375524
Interpolated C_8	15	2	0.487216	0.488982
Interpolated C_8	30	2	0.464799	0.759083
Interpolated C_8	45	1	0.504142	0.504142
Interpolated C_8	90	1	0.555455	0.555455
Steerable C_8, 3×3	All	4	0.175822	0.637367
Steerable C_8, 3×3	15	2	0.310474	0.372270
Steerable C_8, 3×3	30	2	0.175822	0.637367
Steerable C_8, 5×5	All	7	0.225323	0.506883
Steerable C_8, 5×5	8	1	0.367079	0.367079
Steerable C_8, 5×5	15	3	0.331383	0.506883
Steerable C_8, 5×5	30	2	0.225323	0.287781
Steerable C_8, 5×5	45	1	0.246891	0.246891

Table 19: Range of stored post-training 45° relative feature-change diagnostics recovered from original shards. All combines the displayed coverage cells and seeds; it is not a confidence interval. Missing diagnostics are not imputed. Full-precision values and filenames are retained in the audit JSON and per-seed CSV.

test angle	plain target acc. (%)	d_{C_4} (deg.)	exact C_4 T (save)	interp. C_8 T (save)	steer. C_8 T (save)	ideal d_{C_8} (deg.)
105°	73.76	15	15 (165)	15 (165)	15 (165)	15
120°	73.32	30	30 (150)	30 (150)	15 (165)	15
135°	73.51	45	30 (150)	30 (150)	15 (165)	0
150°	73.52	30	30 (150)	30 (150)	15 (165)	15
165°	73.75	15	15 (165)	15 (165)	15 (165)	15

Table 20: Transformation-coverage saving on the complete ImageNet-100 grid (60 runs, three seeds per cell). The target is the same-angle plain model trained at $T = 180$; each entry gives the smallest sampled T whose mean accuracy reaches that target, followed in parentheses by the saved angular band $180 - T$. The table is descriptive on the sampled grid and does not interpolate between coverages. Folded distances are references for the C_4 subgroup and an ideal C_8 action; the interpolated C_8 follows the C_4 coverage requirement in this experiment.

arm	params	profiled GFLOPs/image	peak GiB	train sec/run	interpretation
Plain	11.22M	4.44	0.24	367	plain reference
Exact C_4	11.19M	17.75	0.97	2924	constructed subgroup
Interp. C_8	10.57M	33.92	1.45	4190	approximate generator
Steer. C_8 3×3	7.77M	45.39	1.24	3358	approximate steerable generator

Table 21: Resource controls for the 64px ImageNet-100 coverage family. Parameters and training seconds are means over the 15 runs of each arm. GFLOPs/image counts operations supported by the PyTorch profiler during one forward–backward training step on an RTX 4080; it is not a complete count of all operations. Peak allocated memory uses batch 8 for plain, C_4, and interpolated C_8, and batch 2 for steerable C_8. Memory values therefore describe different batch sizes. These profiles are resource references, not end-to-end hardware or wall-clock equivalence.

target q	exact C_4 T	interp. C_8 T	steer. C_8 T	mean savings (deg.)
72%	15,30,30,30,15	15,30,30,30,15	15,15,15,15,15	156,156,165
73%	15,30,30,30,15	15,30,30,30,15	15,15,15,15,15	156,156,165

Table 22: Sensitivity of the coverage ordering to fixed absolute targets. Entries list the minimum sampled training band at test angles $105/120/135/150/165^\circ$; the final column gives mean savings relative to plain at $T = 180$. The same ordering as the matched plain-reference analysis is obtained at both nearby targets.

test angle	plain T=180 target per-seed target (%)	exact C_4 per-seed first T ($s0/s1/s2$)	interp. C_8 per-seed first T ($s0/s1/s2$)	steer. C_8 per-seed first T ($s0/s1/s2$)
105°	73.61 / 73.72 / 73.95	15 / 15 / 15	15 / 15 / 15	15 / 15 / 15
120°	73.13 / 73.33 / 73.51	30 / 30 / 30	30 / 30 / 30	15 / 15 / 15
135°	73.49 / 73.62 / 73.41	30 / 30 / 30	30 / 30 / 30	15 / 15 / 15
150°	73.45 / 73.54 / 73.56	30 / 30 / 30	30 / 30 / 30	15 / 15 / 15
165°	73.90 / 73.62 / 73.74	15 / 15 / 15	15 / 15 / 15	15 / 15 / 15

Table 23: Secondary seed-level audit of the discrete coverage crossing. For each seed, the target is that seed's own plain T=180 accuracy at the same test angle; entries report the first sampled T reaching that target. The crossing is identical across the three seeds in every arm and angle here. This audit is complementary to the main mean-based estimand (Table 20); it is not a confidence interval for a discrete crossing.

work	studies	measures	domain
Brehmer et al. 2024	hard equiv.	data efficiency at scale	rigid-body / transformer
Adly 2026	hard equiv.	sample complexity rate	synthetic C_n tasks
ours	hard / learned priors	fixed-data coverage, budgets	images: rotation / elastic

Table 24: Complementary questions in the equivariance-vs-augmentation literature. We evaluate fixed-data coverage and learned-prior comparisons under the specified backbones and augmentation controls; these questions overlap with prior work on constructed symmetry.

method	clean	severe	Δ vs plain	Δ vs aug
plain	85.5	78.6	—	—
+ GCLS (ours)	84.8	78.8	+0.1–0.8	−6.7
+ CoordConv	85.6	78.6	+0.06	−6.85
+ DCNv2	85.8	79.3	+0.67	−6.24
+ ORN (one stage)	85.8	79.3	+0.73	−6.17
+ STN	85.8	80.6	+2.02	−4.89
+ TPS-STN	88.3	82.7	+4.14	−2.77
+ augmentation	87.1	**85.5**	+6.91	—

Table 25: Five geometric-prior probes plus GCLS in a common ResNet setup (severe elastic out-of-distribution (OOD) shifts, parameter-matched, 3 seeds). In the listed order, observed gains over plain increase, but each insertion remains below augmentation. These separate-arm comparisons do not measure the benefit of combining a prior with augmentation.

#	Control	Question	Evidence to report
C1	Augmentation and stacked arms	Does the prior add to augmentation?	Prior, augmentation, and their combination on the same task
C2	Hierarchical backbone	Do flat-network gains persist?	Task, architecture, and resource-matched comparisons
C3	Constructed control	How does broad-angle performance compare?	Coverage, per-seed accuracy, and group implementation
C4	Generator audit	Is the claimed action accurately realized?	Angle, input distribution, feature definition, and checkpoint

Table 26: Geometric-Prior Evaluation Protocol (GPEP). Record prospective criteria where available and distinguish them from exploratory follow-ups; report actual seed counts and uncertainty. Internal script records do not constitute external preregistration, and no single control establishes a universal verdict.

prior (suite)	capacity	robustness	reading
soft gauge (rotation, Tiny-IN)	$+\mathbf{5.8}$	$-\mathbf{0.2}$	higher ID accuracy
C_4 / C_8 (rotation, Tiny-IN, hard)	$+3.5$ / $+1.7$	$+\mathbf{2.8}$ / $+\mathbf{3.3}$	reduced OOD drop
$C_4/D_4/C_8$ (rot.-CIFAR, hard)	$-0.8/+1.7/-0.2$	$+\mathbf{24.1}/+\mathbf{23.6}/+\mathbf{25.0}$	reduced OOD drop
soft gauge / DCNv2 (elastic)	≈ 0	$+0.6$ to $+1.0$	small drop reduction
soft gauge (elastic, Blood/Tissue)	-0.5 / -0.1	$+2.0$ / -0.1	aug: $+4.4/+5.7$
STN (elastic)	$+0.2$	$+2.1$	reduced OOD drop

Table 27: Diagnostic decomposition (Eq. 3; Fig. 11). "Capacity" denotes the ID accuracy difference; "robustness" denotes the reduction in OOD drop. These labels do not identify causes or establish augmentation replaceability. The C_8 implementations are approximate; elastic and rotation comparisons concern different tasks.

arm	train	params	id $\pm 30^\circ$	ood $\pm 180^\circ$	90° equiv. err.
plain	$\pm 30^\circ$	286k	97.4	56.7	9.3e-2
ORN (full: ARF N=8 + ORPooling)	$\pm 30^\circ$	289k	95.5	**95.0**	**1.4e-6**
rotaug (plain, full angles)	$\pm 180^\circ$	286k	92.4	92.6	6.3e-2
ORN + rotaug	$\pm 180^\circ$	289k	95.7	95.6	1.4e-6
C_4 G-CNN (hard control, §5)	$\pm 30^\circ$	371k	95.3	94.8	3.9e-6

Table 28: Full-network ORN on rotated-MNIST (§4.4). Under the reported protocol, it exceeds full-angle augmentation by 2.5 percentage points (exploratory paired $p = .002$); the stacked arm exceeds augmentation alone by 3.0 percentage points ($p = .006$). The 90° feature check does not certify exact C_8 equivariance at 45°.

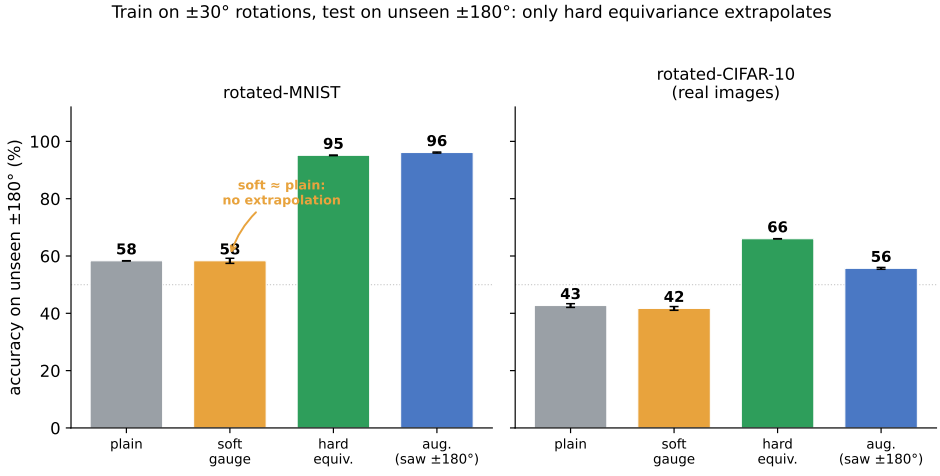


Figure 4: Training on $\pm 30^\circ$ and evaluating over $\pm 180^\circ$: the constructed controls attain substantially higher broad-angle accuracy than their plain baselines, whereas the tested gauge does not (§5). The test band includes both seen and unseen angles; the MNIST hard and soft results use separate experiments.

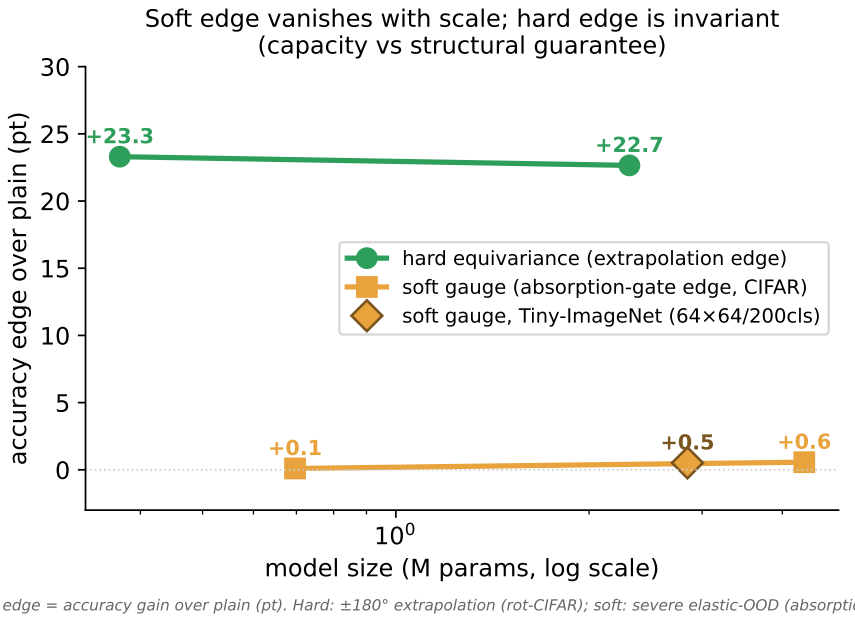


Figure 5: Reported architecture and scale controls (§5). The gauge edge decreases while the constructed control retains a similar advantage at the tested sizes. These comparisons do not establish capacity as the cause or invariance of either trend beyond the measured settings.

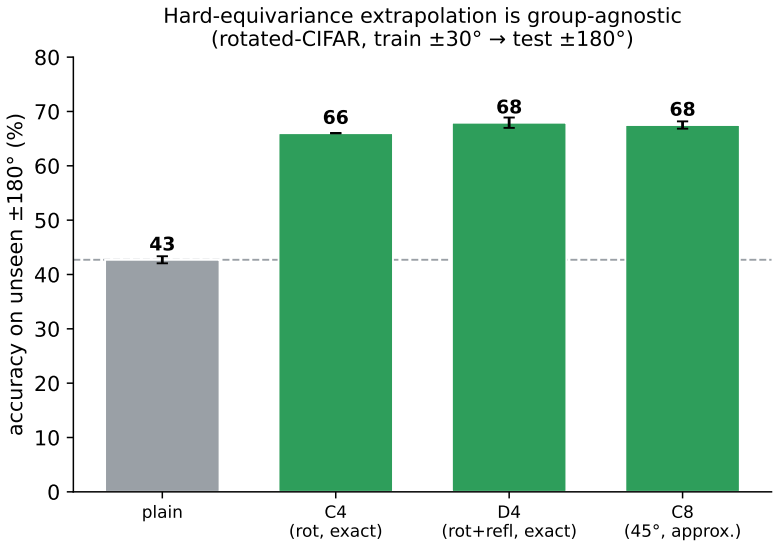


Figure 6: Broad-angle accuracy of the tested C_4, D_4, and approximate C_8 controls (§5). All share a grid-exact C_4 subgroup; this comparison does not establish generality across arbitrary transformation groups.

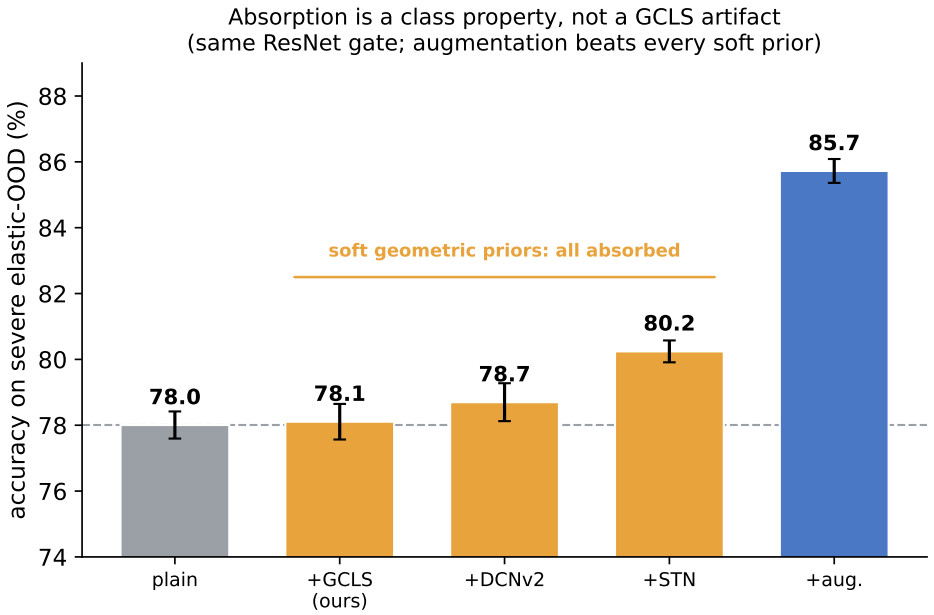


Figure 7: The displayed GCLS, DCNv2, and STN comparisons in the ResNet elastic-OOD experiments (§4.3). GCLS is shown as its within-experiment gain added to the literature-prior experiment's plain baseline; its bar is not a directly measured absolute accuracy on that experiment. The full six-prior comparison is in Table 25.

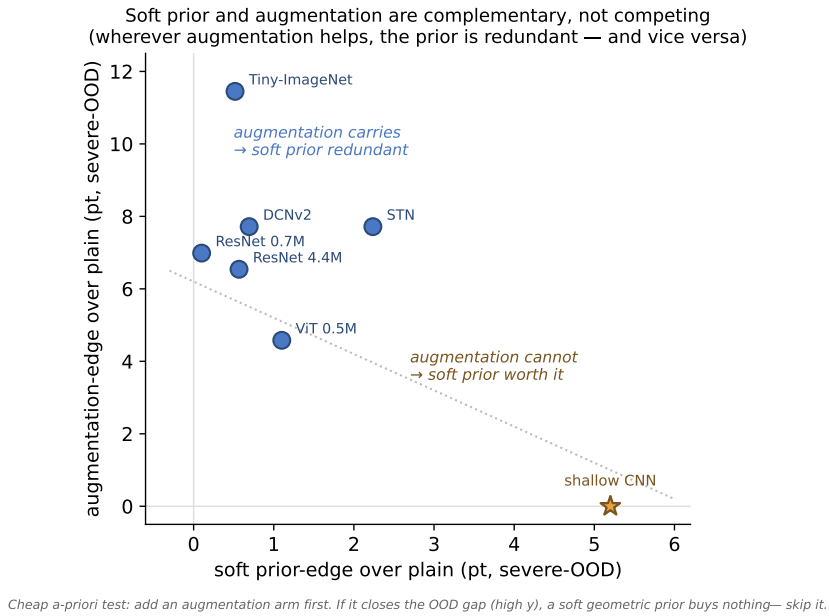


Figure 8: Prior-minus-plain versus augmentation-minus-plain gains in the reported experiments (§7). The hierarchical-backbone cases have smaller prior gains and larger augmentation gains than the displayed flat-network case. These separate-arm results do not measure their combination.

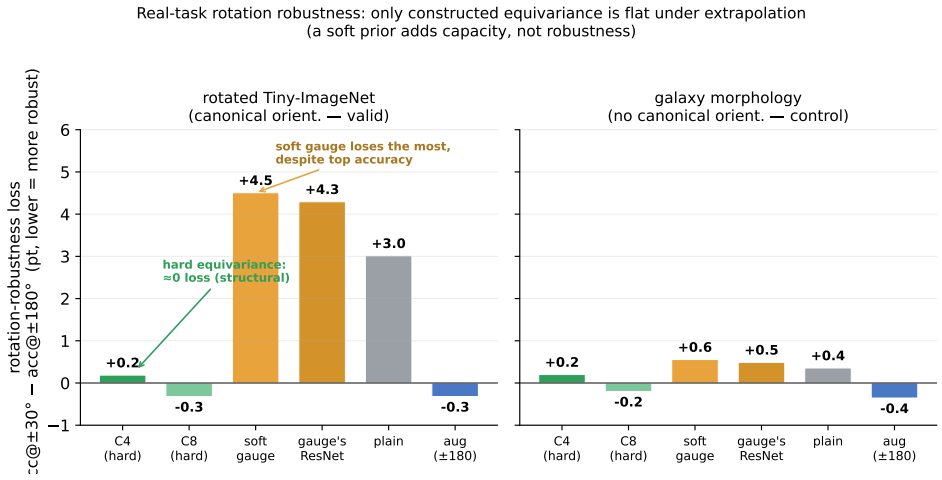


Figure 9: Task-specific rotation penalties (acc@$\pm 30°$ − acc@$\pm 180°$). The C_4 and approximate C_8 controls have small penalties in the reported tasks. Gauge has the largest Tiny-ImageNet drop despite its high starting accuracy; all galaxy-morphology arms change little (§5).



Figure 10: Measured degradation curves with narrow-band training and test bands extending from $30°$ to $180°$. Constructed controls vary little compared with plain and gauge on Tiny-ImageNet; all galaxy arms vary little. These are observations at the sampled bands, not coverage-independent guarantees (§5).

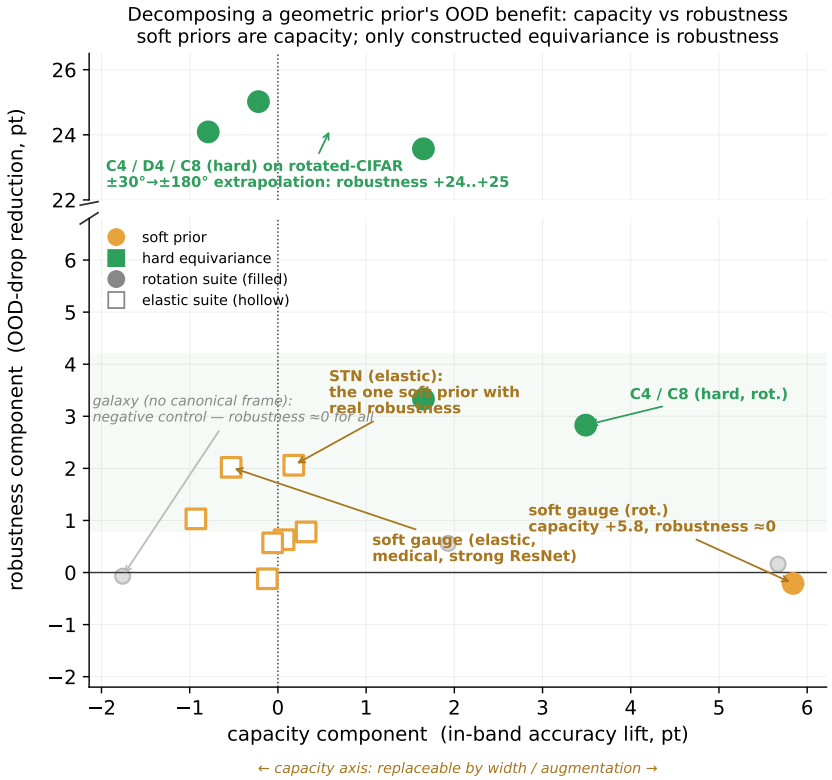


Figure 11: Diagnostic in-distribution gain versus reduction in OOD drop (Table 27); the broken y-axis includes the rotated-CIFAR constructed controls. The "capacity" label is descriptive and does not identify a mechanism (§6).

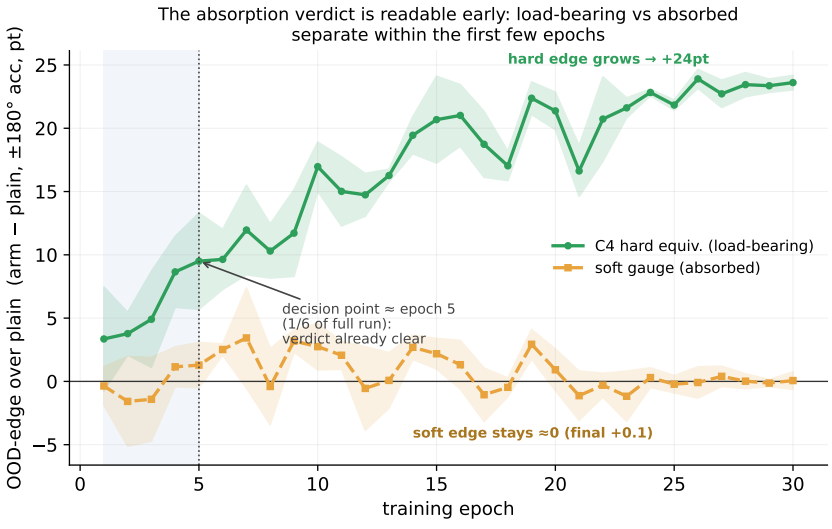


Figure 12: Per-epoch OOD gain over plain in the early-signal probe: the constructed arm separates from epoch 1 while the measured gauge gain remains near zero. This single probe does not validate early prediction of a new method's final benefit (§7).

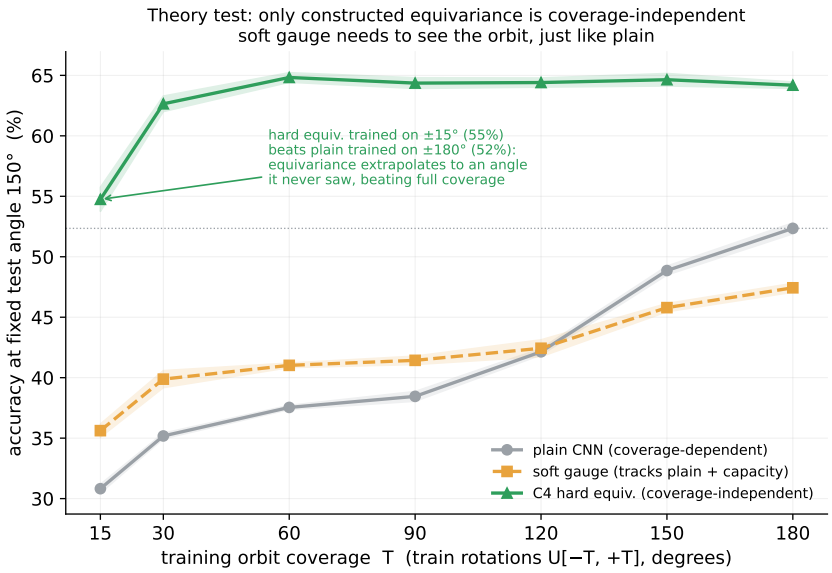


Figure 13: The CIFAR coverage sweep in full (§3): accuracy at fixed 150° vs training coverage.