THE MEMORIZATION COST OF GENERALIZATION: A QUANTITATIVE CAUSAL LAW IN NOISY-TRAINED DEEP NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep networks trained on noisy labels do two things at once: they learn generalization from the clean samples and they account for the noise samples by memorizing them. We show that these two processes occupy complementary channels of the representation's topology. Using a two-level persistence pipeline (Topo^2) that separates H_1 persistent homology into a within-class component H_1^{within} (loops internal to a single class, the generalization channel) and a cross-class component H_1^{cross} (loops that close only across classes, the memorization channel), we establish that memory and generalization are causally separable: memory can be added, stripped, and — the central result — its cost on generalization can be billed quantitatively. Along the overlay path (freeze a clean-trained model, then unfreeze K flipped samples with clean samples still present), the memorization cost obeys the law

$$\text{va} \;=\; \text{va}_{\text{FM0}} - C\,\frac{\eta}{1-\eta}\,\text{mem},$$

i.e. each 1% of memorized flipped samples costs $C \cdot \frac{\eta}{1-\eta}\%$ validation accuracy, with the slope proportional to the noise:clean sample ratio $\eta/(1-\eta)$. At the shared ∼11M reference capacity, the slope constant clusters near 0.38 across datasets and class counts: $C = 0.386$ (CIFAR-10), 0.397 (SVHN), 0.384 (CIFAR-100, 100 classes), and 0.366 (VGG on SVHN); the per-setting spread is the $\approx \pm 0.02$ image term of the displacement mechanism, not a capacity signature. The law also holds on *real human label noise*: on CIFAR-10N the calibrated constant predicts the measured memorization slope to a mean ratio of 1.04 ± 0.02 (3 seeds), while maximally class-confusable flips bound the coefficient's noise-structure sensitivity at $\approx +32\%$. Across capacities C varies systematically ($0.47 \rightarrow 0.34$ over a width sweep) and is traced to a normalized clean-sample feature displacement, $C \approx 0.5\, d_{\text{clean}}$ (Zhang et al., 2026). Every calibration cell is 3-seed. The constant is a through-origin fit; a low-noise extrapolation to $\eta = 20\%$ stays within 14% on both datasets but with larger residual scatter, delimiting the law's cleanest domain to the mid-noise range $\eta \in [40, 60]\%$. Memory, we have shown, is an invertible, stackable layer, not an uninterpretable black-box capacity: its cost to generalization obeys a quantitative causal law with a calibrated constant that clusters near 0.38 across datasets and class counts at the shared reference capacity, with its capacity dependence mechanistically traced.

1 INTRODUCTION

Motivation. Deep networks can fit arbitrary noisy labels to zero training error (Zhang et al., 2017), yet this tells us nothing about how generalization coexists with memorization. In the label-noise literature, memorization is treated either as a form of overfitting to be avoided, or as a capacity with unknown cost. Neither treatment explains the basic observation that noisy-trained networks both memorize and generalize at the same time.

Our question. Are memorization and generalization *separable things* in the geometry of the representation? And if they are, can we *causally* add memory, *strip* it, and *bill* its cost on generalization

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.

quantitatively? The label-noise setting is the natural laboratory for this question, because it turns memorization from an accident into a controlled variable: the fraction η of flipped samples is under our direct control, and so is the amount of memorization K via the overlay protocol (Sec. 3).

Three contributions.

1. **(C1) Geometric separability.** Memorization and generalization are causally separable topological channels: H_1^{cross} (cross-class rings) is the memorization account, H_1^{within} (within-class rings) is the generalization account. Memory is an *invertible layer* addable while clean samples are present; stripping it restores near-ceiling generalization (va $0.63 \rightarrow 0.92$, close to the FM0 ceiling).
2. **(C2) A quantitative law.** The memorization cost of generalization is $\text{va} = \text{va}_{\text{FM0}} - C\,\frac{\eta}{1-\eta}\,\text{mem}$, with a slope constant that clusters near $C \approx 0.38$ at the ∼11M reference capacity — across datasets (CIFAR-10 0.386, SVHN 0.397) and class counts (CIFAR-100 0.384), spanning 0.366 (VGG on SVHN) to 0.397; the per-setting spread is the $\approx \pm 0.02$ image term, not a capacity signature, and C is capacity-dependent in general (Sec. 5.6).
3. **(C3) A mechanism model.** The slope is the decision-boundary displacement ratio $\eta/(1-\eta)$: noise samples push the decision boundary the wrong way, clean samples pull it back, and a memorized sample hurts in proportion to the noise:clean concentration. The class-count independence judged by CIFAR-100 shows the cost is set by the *sample concentration ratio*, not by how many distinct classes the flipped labels occupy; the calibrated constant transfers to real human annotation error (CIFAR-10N), and flip structure perturbs it only weakly (Sec. 4.3).

Relation to the memorization literature. The memorization-theory literature asks *which* samples are memorized and *why* the long tail needs them (Feldman, 2020; Feldman & Zhang, 2020; Arpit et al., 2017). We ask a complementary question: *where* in the geometry memorization lives and *how much* generalization it costs. The two questions are orthogonal; our law answers the second.

2 Background and Definitions

2.1 Representation geometry and persistent homology

We study the topology of a network's penultimate-layer features. Given a point cloud of feature vectors, persistent homology computes the birth and death of topological features (connected components, loops, voids) across a scale filtration. We use H_1 — the loop count — because loops capture the connectivity structure of local neighborhoods that global statistics cannot (established by the exclusion chain of Zhang et al. (2026)). The pipeline builds an interpolation graph between labeled samples and measures the rank-1 persistent homology of local neighborhoods; the exact construction is described in Sec. 3.

2.2 Two topological channels

The pipeline separates H_1 into two components:

- H_1^{within} — loops internal to a single class, the structure within a class manifold. This is the *generalization account*: it is a function of the training stop position (Sec. 3), not of the noise exposure per se.
- H_1^{cross} — loops that close only across classes, linking different class neighborhoods. This is the *memorization account*: it scales monotonically with injected memory under causal intervention (Sec. 4).

As a rule, absolute Betti values are pipeline-relative normalization quantities (depending on the number of interpolation pairs, etc.; Sec. 3). All $H_1^{\text{within}}/H_1^{\text{cross}}$ numbers in this paper are comparable within the fixed pipeline; we make no cross-pipeline absolute claims.

2.3 THE FM0 PRESCRIPTION

The **FM0** prescription freezes (zeros the loss on) flipped samples from epoch 0, so the model is trained on clean samples only. FM0 reaches each setting's generalization ceiling and is the natural "no-memory" anchor: mem ≈ 0.004 (never memorizes), va at the setting's ceiling. The data-driven FM0-hat (mask learned from data) recovers 90–97% of the FM0 gain and is deployable, but FM0 itself is the mechanistic anchor used throughout this paper.

2.4 THE TLS FRAMEWORK

The framework that organizes the results is **TLS** (topological layering): memory is *anchored* to generalization (silencing the clean signal collapses the representation, va $0.92 \rightarrow 0.04$), memory is *addable* on top of a live generalization channel, and the cost of memory is *quantitative* — the law of Sec. 4.

3 METHOD

3.1 PROTOCOL

CIFAR-10 and SVHN are trained with ResNet-18 (and VGG for the architecture axis), 200–300 epochs with the standard schedules; the flip mask is regenerated deterministically from the seed, so the noise is bit-identical across all runs of a setting. Two model families (ResNet-18 / VGG) cover the architecture axis. Seeds: every calibration cell is 3-seed — CIFAR-10 $\eta \in \{30, 40, 50, 60\}$, SVHN $\eta \in \{30, 40, 50, 60\}$, CIFAR-100 $\eta \in \{30, 50\}$, and VGG $\eta \in \{30, 50, 60\}$; the out-of-grid low-noise probes (η=20, CIFAR-10 and SVHN) are 3-seed as well (Sec. 4).

3.2 THE H_1 PIPELINE AND ITS NORMALIZATION

The Topo^2 pipeline computes a cross-class decomposition: for each class pair and interpolation position it builds a local neighborhood graph and measures rank-1 persistent homology, then separates rings into the within-class and cross-class components of Sec. 2.2. Two normalization facts are load-bearing for the honest reading of every number in this paper:

- **Pipeline-relative absolute values.** The cross absolute count is proportional to $N_{\text{pairs}} \times N_{\text{interp}}$; a reported cross value is only meaningful against the pipeline that produced it. We report pipeline specification with every number.
- **RNG sampling noise.** Cross readouts carry ±15–30% RNG sampling noise; fine cross-model comparisons include error bars.

3.3 DUAL-SCOPE MEMORY READOUT

Memory is measured as $\text{mem}_{\text{noisy}}$, the fraction of flipped samples whose prediction equals their noisy label. This readout dichotomizes cleanly across 9/9 setting-seed combinations: CONT (co-trained on noisy labels) memorizes essentially all flipped samples (0.999), FM0 memorizes essentially none (0.004).

3.4 THE OVERLAY PROTOCOL (CAUSAL INTERPOLATION)

1. Train a clean model (FM0): flipped samples' loss is zeroed from epoch 0. This is the generalization anchor, va_{FM0}.
2. Freeze FM0, then unfreeze the loss on K flipped samples while clean samples remain present. This causally injects memory K into the frozen generalization channel — an interpolation from FM0 (K=0) to co-trained (CONT, K= all).
3. Measure va and mem at each K. The curve va vs. mem is the overlay cost curve.

Because the generalization channel is frozen, any va drop along the curve is **causally attributable** to the injected memory, not to training-time confounds.

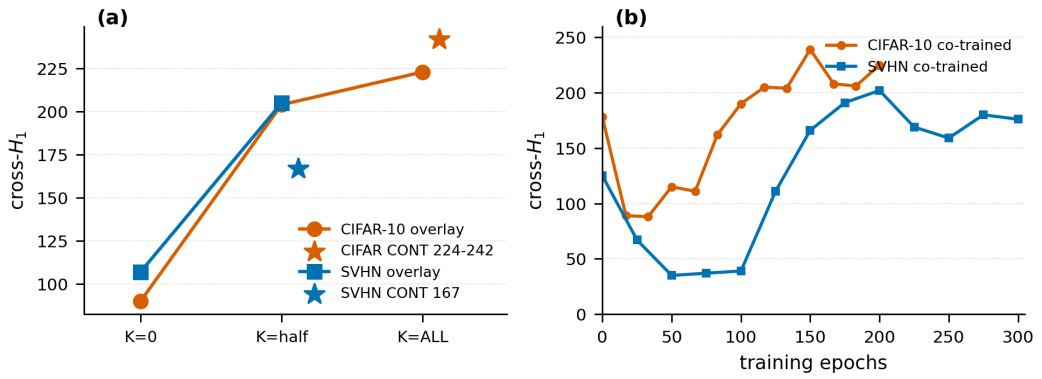


Figure 1: **Cross-H_1 is the geometric signature of memory.** (a) overlay: H_1^{cross} scales monotonically with injected memory on both datasets; (b) cross trajectory: the memorization signature rises with memorization onset and settles at the converged value (SVHN peak 202 at ep200, settled 176; CIFAR peak 239 at ep150, settled 225).

3.5 Calibration estimator

The law's slope k at each noise level is the two-point slope between FM0 and the half-memory point,

$$k = \frac{\text{va}_{\text{FM0}} - \text{va}_{K=\text{half}}}{\text{mem}_{K=\text{half}} - \text{mem}_{\text{FM0}}},$$

which agrees with the full-curve fit (difference <1%). The constant C is the through-origin slope of k against $\eta/(1-\eta)$, $C = \frac{\sum x_i k_i}{\sum x_i^2}$ with $x_i = \eta_i/(1-\eta_i)$, over the $\eta \in \{30, 40, 50, 60\}$ grid.

4 Results

4.1 (C1) Memory is separable: FM0 vs. CONT

The dual-scope readout dichotomizes 9/9 setting-seed combinations (Fig. 4, App. D): CONT memorizes the flipped samples (mem $\approx$ 0.999), FM0 does not (mem $\approx$ 0.004). The two checkpoints are opposite ends of the same protocol axis, not two unrelated models. Memory is *anchored* to generalization: silencing the clean signal collapses the representation (va 0.92 $\rightarrow$ 0.04).

4.2 (C1) Memory is addable: causal overlay

Unfreezing K flipped samples on top of the frozen FM0 anchor makes H_1^{cross} scale monotonically with injected memory in 3/3 seeds: CIFAR-10 H_1^{cross} 90 $\rightarrow$ 204 $\rightarrow$ 223 (CONT 224–242); SVHN 107 $\rightarrow$ 205, overshooting the CONT reference (167); the overshoot lies within the ±15–30% RNG band on absolute cross values (Sec. 3). H_1^{within} moves only within training-volume artifacts. The causal channel is memory, not training time: the generalization channel is frozen, so the monotone cross rise is attributable to injected memory (Fig. 1).

4.3 (C2) The memorization cost law

Along the overlay path, va falls linearly in mem (Fig. 2):

$$\text{va} = \text{va}_{\text{FM0}} - k\ \text{mem}, \qquad k = C\,\frac{\eta}{1-\eta}.$$

The calibration table (through-origin fit over the two-point slopes) is Table 1; the fitted constants are C = **CIFAR-10** 0.386 (R^2=0.985) / **SVHN** 0.397 (R^2=0.998) / **CIFAR-100** 0.384 / **VGG** 0.366. The four settings cluster near 0.38 (range 0.366–0.397): the per-setting offsets are the $\approx$ ±0.02 image term of the displacement mechanism (Sec. 5.6), and the controlled class-count pair (CIFAR-10 vs. CIFAR-100) agrees to 1%. Every calibration cell is 3-seed; extending the CIFAR-10 η=50 cell

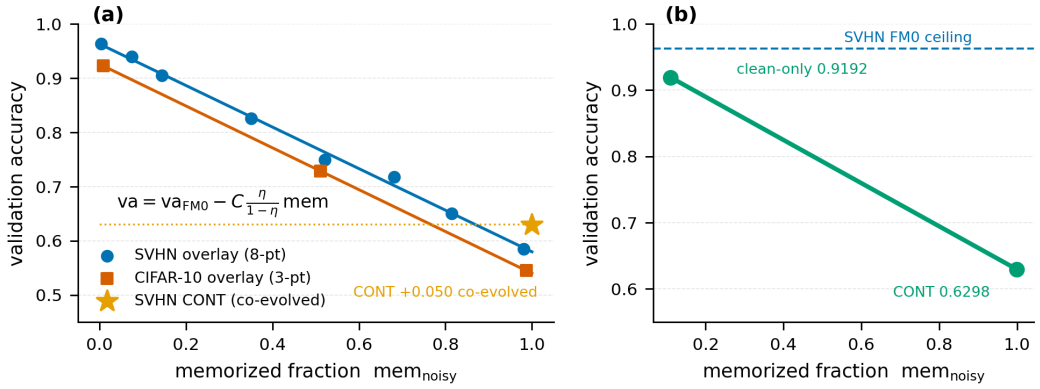


Figure 2: **Memory is an invertible layer.** (a) Overlay cost curve: va = 0.963 − 0.383 mem on SVHN (8-pt), va = 0.925 − 0.386 mem on CIFAR-10 (3-pt); CONT sits above the curve on SVHN (+0.050, co-evolution advantage). (b) Stripping memory from CONT (clean-only continue) restores va 0.63 → 0.92, near the FM0 ceiling.

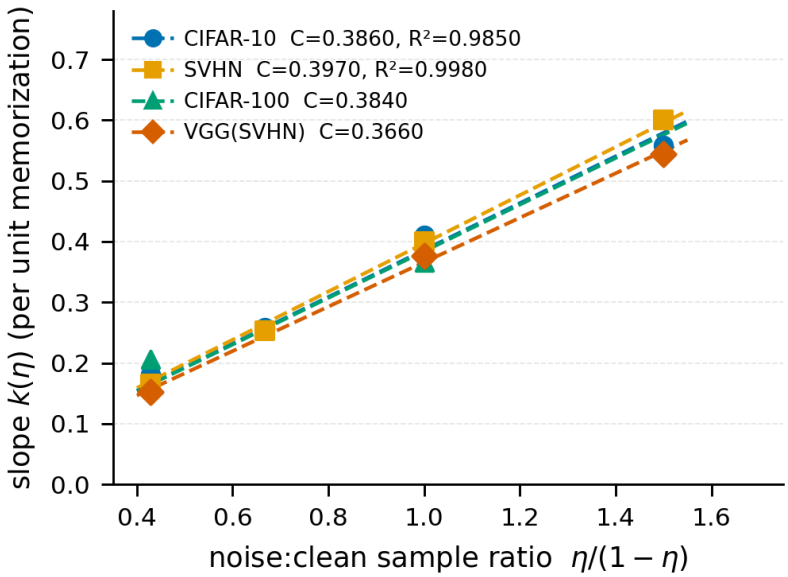


Figure 3: **The slope k is linear in $\eta/(1-\eta)$ across four settings** (through-origin, over the calibrated mid-noise grid; low-noise residuals in Fig. 6), with fitted constants $C = 0.386$ (CIFAR-10), 0.397 (SVHN), 0.384 (CIFAR-100), 0.366 (VGG).

from one seed to three (k=0.410 ± 0.023) refines the CIFAR-10 constant from its single-seed value 0.3801 to 0.386 and leaves every judgment below unchanged; the per-level slopes, seed counts and CVs are tabulated in Table 1 (App. A). On the same four-level grid the $\eta/(1-\eta)$ and η^2 forms remain statistically indistinguishable (Fig. 5): linear R^2=0.981, power-law 0.991, η^2 0.996, $\eta/(1-\eta)$ 0.985 — we retain $\eta/(1-\eta)$ because it is the mechanistic form, the noise:clean ratio of the boundary-displacement model (Sec. 5). Caliber: the law is calibrated *across converged overlay endpoints* (varying K); k is the FM0→endpoint chord slope. Along a single trajectory the va–mem path is concave, so the linear form describes the endpoint family, not the path.

4.3.b CIFAR-100: the class-count judgment. The decisive test of whether C is set by sample concentration or by the class distribution is the 100-class setting. CIFAR-100 $\eta = 50$ and $\eta = 30$, 3 seeds each: $k = 0.366$ and $k = 0.206 \pm 0.006$ (CV 2.9%). Through-origin fit: $C = 0.384$, vs. CIFAR-

10 0.386 — within 1%. **Judgment: at fixed capacity, C is independent of the number of classes; the law generalizes across class counts**, and the independence is pinned at matched image content by retraining the same images on 10 superclass labels (C=0.370 vs. fine 0.369; Sec. 5.6). Honest boundary: the η=30 single-point bias is +25% here (Fig. 6), and the η=50 cell is the calibration record of Zhang et al. (2026) (0.366; a rerun for the Sec. 5.6 displacement arc read 0.3714, within the cell's run-to-run variance).

4.3.c VGG: the architecture axis. VGG on SVHN, 3 seeds: η=50 gives k = 0.397/0.370/0.362, mean 0.376 ± 0.015 (CV 3.9%, comparable to the ResNet cells); through-origin C = 0.366 over η=$\{30, 50, 60\}$. **Judgment: the architecture changes the constant within the cluster (VGG** 0.366 **at its low end, vs. ResNet-on-SVHN** 0.397**) while preserving the $\eta/(1-\eta)$ function form — the law's shape is architecture-independent.** The offset is consistent with the capacity axis (14.7M vs. ∼ 11M predicts a lower C on the width trend); architecture and capacity are not separately identified at a single capacity, which the matched-capacity anchors of Sec. 5.6 address.

4.3.d The noise-structure axis: real human and structured flips. Every cell above is calibrated on uniform random flips. Two further noise structures, at the same CIFAR-10 ResNet capacity, test whether C is noise-type invariant. *Real human annotation error.* CIFAR-10N (Wei et al., 2022) is a set of human-annotated CIFAR-10 labels; the *worse* split carries η=0.402 measured flips. Running the identical FM0→half overlay under the dataset's own human mask gives a two-point slope k=0.271±0.004 (3/3 seeds), versus the calibrated prediction $C\,\eta/(1-\eta)$=0.2597 (unrounded CIFAR-10 constant 0.3863) — ratio 1.043 ± 0.017. **The calibrated constant transfers to real human annotation error.** Both checkpoints share the mask, so the retained-subset caliber of the FM0 anchor ($\mathrm{va}_{\mathrm{FM0}}$=0.915) cancels in the slope difference. The geometric signature transfers as well: under real flips H_1^{cross} rises from an FM0 mean of 106 to a co-trained CONT mean of 184, with no seed overlap (min CONT 178, max FM0 121). *Structured flips.* CIFAR-10 asymmetric (confusion-pair) noise places every flip on a nearest-neighbor confusion pair (0↔2, 1↔9, 3↔5, 4↔7; classes 6 and 8 are never flipped). At η=40% the same protocol gives k=0.340 ± 0.014 versus the prediction 0.2575 — ratio 1.320 ± 0.055, all three seeds above the law. **Judgment: the** $\mathrm{va}=\mathrm{va}_{\mathrm{FM0}} - C\,\frac{\eta}{1-\eta}\,\mathrm{mem}$ **law form survives all three noise structures; the effective coefficient carries a weak dependence on how tightly flips align with class structure** (uniform ≈ C_{ref}, real human error 1.04×, maximally class-confusable flips 1.32×), the ordering predicted by the orthogonal-carving mechanism of Sec. 5.6. The structured-flip deviation is measured at the converged endpoint, the calibration's own convention, and is a structure-sensitivity bound rather than a lower bound: the overlay arc is not monotone (memory saturates early while va partially recovers), so intermediate-time comparisons would overstate it.

4.4 (C1) Memory is invertible: the strip operation

Continuing CONT on clean-only data removes memory (mem 0.999 → 0.11) and restores generalization (va 0.63 → 0.92), close to the FM0 ceiling (Fig. 2b). The residual gap to FM0 (0.92 vs. 0.96) is noise-gradient residue on the generalization channel; the memory layer itself is fully invertible. Invertibility does not decay with memory age: zeroing flipped-sample loss at epoch 200/225/250 of an SVHN η=40 run and continuing at restored lr (0.01, 100 ep) strips even saturated memory (mem 0.905 → 0.058; va flat at 0.938–0.943 across all three ages), while the same intervention under the schedule's decayed tail leaves mem at 0.42–0.91: recovery is bounded by the remaining learning-rate budget, not by memory age.

4.5 (C3) Co-evolution: CONT sits above the overlay curve

Co-trained CONT achieves the same memory load at strictly lower generalization cost than the post-hoc overlay: SVHN-ResNet +0.050, CIFAR-10 +0.006, VGG +0.013. The effect is SVHN-ResNet specific; its determining factor remains open (Sec. 7).

5 MECHANISM: THE DECISION-BOUNDARY DISPLACEMENT RATIO

5.1 GEOMETRIC INTUITION

A noise sample pushes the decision boundary the wrong way; a clean sample pulls it back toward the correct side. When a fraction mem of the training signal is memorized noise, the displacement of the boundary at the population level is in proportion to the ratio of noise to clean concentration. The model can protect generalization only by keeping clean samples' influence dominant, and the marginal cost of each memorized sample scales with $\eta/(1-\eta)$.

5.2 DERIVATION OF $k = C\,\eta/(1-\eta)$

Let C be the memorization cost coefficient per single noise sample (the va loss contributed by one memorized flipped sample relative to one clean sample's generalization contribution). Then over the training distribution the boundary displacement ratio is

$$\frac{\text{noise concentration}}{\text{clean concentration}} = \frac{\eta}{1-\eta},$$

so the slope of the cost curve is $k = C \cdot \eta/(1-\eta)$. The measurement protocol (two-point FM0→half, Sec. 3) estimates k directly; fitting k vs. $\eta/(1-\eta)$ gives C.

5.3 SELF-CONSISTENCY OF $C \approx 0.38$

$C \approx 0.38$ equals the full-memory cost at η=50%, $\text{va}_{\text{FM0}} - \text{va}_{\text{CONT}} = 0.379$ on CIFAR-10: the per-unit slope recovered from the two-point protocol is consistent with the end-to-end cost of fully memorized noise — the constant is not an artifact of the slope estimator. On SVHN the gap is smaller, 0.333, because CONT sits *above* the overlay curve there (+0.050 co-evolution advantage, Sec. 4.5); self-consistency is established on the setting that is not co-evolution-lifted, and the lift itself is the §4.5 effect.

5.4 LOW-NOISE EXTRAPOLATION AND RESIDUALS ($\eta = 20\%$)

$\eta = 20\%$ is an extrapolation 42% below the calibrated grid, and it is the one regime where 3-seed replication changed the picture. Both settings stay close to the line: SVHN gives k=0.092 ± 0.005 vs. predicted 0.099 (ratio 0.93), CIFAR-10 k=0.110 ± 0.014 vs. predicted 0.097 (ratio 1.14). Adding η=20 leaves the constant essentially unmoved (5-point through-origin C=0.387 / 0.397, both within 0.3% of the in-grid fit): the low-noise point has low leverage. The residual is noisier here, though — CIFAR-10's per-seed ratios span 1.29/1.16/0.95 (CV 12.6% vs. 2–6% in-grid), straddling the law, so the apparent +31% bend of the single-seed estimate was largely seed selection. The surviving statement is weaker: the first-order form is intact but its scatter grows, so we delimit the law's cleanest domain to the mid-noise range rather than claim a dataset-dependent curvature. Fig. 6 (App. D) shows both the η=30 single-point systematic bias and the η=20 extrapolation.

5.5 JUDGMENT: THE CONCENTRATION MODEL SURVIVES, THE LABEL-SPREAD MODEL DOES NOT

The original prediction — that the non-10-class dataset would show a departure from $C = 0.38$ (a concentration model of $\eta/(1-\eta)$) — was *not* supported: CIFAR-100 gives $C = 0.384 \approx 0.38$ (3-seed at both η=30 and η=50). The mechanism is therefore that C is set by the sample concentration ratio $\eta/(1-\eta)$ (established), and does *not* depend on how many distinct classes the flipped labels occupy (spread over 99 vs. 9 classes, C unchanged). Flip *structure* is a separate, weaker axis: confusion-pair flips raise the per-memory cost by 32% and real human flips track the uniform calibration at mean ratio 1.04 (Sec. 4.3). The low-noise extrapolation is first-order with growing residual scatter (no longer read as a dataset-specific curvature, Sec. 5.4). The mechanism is still converging; the law is currently a *phenomenological* law (strong empirical regularity).

5.6 THE CAPACITY DEPENDENCE OF C, AND ITS MECHANISTIC REDUCTION

The four-setting agreement near 0.38 is a reference-capacity value, not a capacity-independent constant. A CIFAR-10 width sweep (3-seed mean) moves C monotonically, 0.471 $\rightarrow$ 0.342 ($w0.5 \rightarrow w2.0$, Fig. 7a); the four settings of Sec. 4 agree because they share a ∼11M reference capacity (Zhang et al., 2026). Across capacities C varies systematically; fixed capacity is a stated validity condition of the law.

The dependence is mechanistically traced by Zhang et al. (2026): at fixed image family, $C \approx 0.5\, d_{\text{clean}}$ (through origin, R^2=0.92, Fig. 7b), where d_{clean} is the normalized clean-sample feature displacement during overlay; a head/backbone swap shows the cost is 100% feature-mediated. Class-count independence at fixed capacity is verified by *retraining* on 10 superclass labels (a coarse merge of the same CIFAR-100 images): C=0.370 vs. the fine C=0.369 (η=50 / K=12500 caliber). At that caliber the residual image term is $\approx \pm 0.02$.

How local is the reduction to the width family? We measured (C, d_{clean}) at η=50 / K=12500 (3-seed mean, CIFAR-10) for three further architectures — WRN28-10 (36.5M parameters), VGG16 (14.7M), and an attention-hybrid (a ResNet18 backbone with two transformer layers, 14.9M) — and asked whether each falls on the width-family line (Fig. 7b). A wide residual network is *absorbed*: C=0.297 (sd 0.006), d_{clean}=0.572, giving C/d_{clean}=0.520, within ±5% of the family slope. This explains its apparent offset below the width trend in parameter space (Fig. 7a): at larger residual capacity the orthogonal-carving projection is stronger, so the clean displacement d_{clean} itself shrinks and C follows. At *matched* capacity the relation does not transfer across families: at ∼14.7M parameters the three architectures give ResNet C=0.369 / hybrid C=0.423 / VGG C=0.443 while their displacements order the opposite way (d_{clean}, 0.777 / 0.699 / 0.700; family d_{clean} at the s42 reference, anchors 3-seed). VGG16 and the hybrid sit +28% and +22% above the width-family line (C/d_{clean}=0.633 and 0.604).

Honest boundary: d_{clean} compares displacement in architecture-specific representation spaces; mean-$\|\cdot\|$ normalization removes scale but not geometry, so part of the +22–28% residual may be a feature-space incomparability artifact rather than a mechanism gap. The conservative reading: d_{clean} is a verified mediator *within* the residual-convolutional family (across width and width×depth); across families it is necessary but not sufficient, and parameters remain the coarser cross-family attribution.

This upgrades the law from a mysterious shared constant to a coefficient with stated validity conditions (fixed capacity *and* residual-conv family) and a traced origin (clean-sample displacement); the remaining open items are the ±0.02 image term and the +22–28% cross-family residual (Fig. 7, App. D).

6 RELATED WORK

Memorization theory. Zhang et al. (2017) established that networks can memorize arbitrarily noisy labels; Arpit et al. (2017) asked when memorization happens; Feldman (2020) proved that long-tail generalization can *require* memorization; Feldman & Zhang (2020) identified *which* samples are memorized via influence estimation. We add the complementary question: which geometric channel memory occupies, and how much generalization it costs.

Learning with noisy labels. The practical literature trades off against memorization: Co-teaching (Han et al., 2018), early-learning regularization (ELR) (Liu et al., 2020), and DivideMix (Li et al., 2020). On CIFAR-10 η=40 the ordering is DivideMix 0.94–0.95 > FM0 oracle 0.92 > ELR+ 0.91 > ELR 0.89 > FM0-hat 0.85–0.87 > Co-teaching 0.80–0.84. We state this plainly: **FM0 is not SOTA** (FM0-hat is mid-tier). The value of FM0 is mechanistic — a clean ceiling and a unified language for the memorization cost — not method competitiveness.

Representation geometry / topological data analysis. Naitzat et al. (2020) is the PH-in-DL precedent (topology evolves with training). We follow the survey literature (Ballester et al., 2023); see also Pun et al. (2022) and Hensel et al. (2021) for surveys of persistent-homology-based learning.

Generalization theory. Beyond norm/width/margin bounds, we add a new dimension: the *geometric cost* of memorizing noise samples, measured as a causally attributable topological quantity.

7 Discussion and limitations

Cross-H_1 is a geometric signature, not a deployed predictor. Partial correlations $\mathrm{corr}(H_1^{\mathrm{cross}}, \mathrm{va} \mid \mathrm{mem})$, n=37: CIFAR −0.45 to −0.56 (ΔR^2+0.046), SVHN +0.08 (ΔR^2+0.000), pooled −0.39. Weak and setting-dependent — cross-H_1 is the geometric signature of memory, not a deployable generalization predictor. The causal scaling of Sec. 4 (monotone under intervention) is a different claim and is unaffected.

Scope of the law. Validity conditions, all stated rather than omitted: (i) *capacity* — the constant is a reference-capacity value; across capacities it moves systematically and is traced to clean-sample displacement (Sec. 5.6); (ii) *noise structure* — the form is unchanged but the coefficient is weakly structure-dependent, from real human error (predicted at mean ratio 1.04) to maximally class-confusable flips (+32%; Sec. 4.3), the ordering matching the orthogonal-carving mechanism; (iii) *noise range* — the clean domain is $\eta \in [40, 60]\%$; at η=20% the extrapolation stays within 14% but per-seed scatter grows (Sec. 5.4), $\eta/(1-\eta)$ being first-order. The mechanism's detailed form remains the paper's weakest link (Sec. 5.5); the capacity reduction closes part of this gap.

The coefficient is an endpoint quantity. C is defined on converged overlay endpoints; the trajectory itself is concave (early memory costs more per unit), so C inherits the endpoint convention of Sec. 3; the trajectory's own characterization is future work.

The saturation endpoint. Overlaying to the CONT load reaches CONT-level cost, with a two-factor gap: the memory tail not fully saturated plus the SVHN-ResNet co-evolution advantage (Sec. 4).

Protocol applicability, not law applicability. The overlay protocol needs an FM0 anchor that actually reaches its setting's clean ceiling, and two architectures fail that precondition under our recipe budget: a ViT-Small anchor does not enter the memorization regime at all (mem ≈ 0.05–0.11 at the overlay endpoint, versus 0.51 for ResNet-18 in the same protocol), and a ConvNeXt anchor is undertrained ($\mathrm{va}_{\mathrm{FM0}}$=0.586, single seed), leaving C unmeasurable in both. These are boundaries of the *measurement protocol* — a learnable FM0 baseline is a stated precondition — not evidence against the law; we report them rather than selecting the architectures that cooperate.

Methodological honesty. Absolute H_1^{cross} values are pipeline-relative (Sec. 3); every calibration cell is 3-seed (Table 1), with only the out-of-grid η=20 probes carrying larger per-seed scatter. Training and readout use an *oracle noise mask*: the flipped samples and their noisy labels are known to the experimenter (bit-identical across seeds, App. A), both to build the FM0/CONT protocols and to measure $\mathrm{mem}_{\mathrm{noisy}}$. The law therefore characterizes memorization under a known mask, not label-noise detection without one. For the real-noise test (Sec. 4.3) the mask is not an assumption but the dataset's own annotation: CIFAR-10N's human labels differ from the clean labels on exactly the flipped set we use.

8 Conclusion

Memorization and generalization are separable in the representation's topology (a fact); causally addable and strippable (a mechanism); and their cost is quantitatively billable (the law):

$$\mathrm{va} \;=\; \mathrm{va}_{\mathrm{FM0}} - C\,\frac{\eta}{1-\eta}\,\mathrm{mem}.$$

The loop from intuition to experiment to law is closed: the constant is invariant across datasets and class counts at the reference capacity (range 0.366–0.397), traced across capacities to clean-sample feature displacement (Sec. 5.6), and transfers to real human annotation error (CIFAR-10N, ratio 1.043; Sec. 4.3). The residual image term (±0.02) remains open; the law stands as a phenomenological law (Sec. 5.5).

REFERENCES

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research (PMLR)*, pp. 233–242, 2017.

Rubén Ballester, Carles Casacuberta, and Sergio Escalera. Topological data analysis for neural network analysis: A comprehensive survey. *arXiv preprint arXiv:2312.05840*, 2023.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 954–959, 2020. doi: 10.1145/3357713.3384290.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 2881–2891, 2020.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 8536–8546, 2018.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021. doi: 10.3389/frai.2021.681108.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.

Chi Seng Pun, Si Xian Lee, and Kelin Xia. Persistent-homology-based machine learning: a survey and a comparative study. *Artificial Intelligence Review*, 55(7):5169–5213, 2022. doi: 10.1007/s10462-022-10146-z.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations (ICLR)*, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Zhanbo Zhang, Ming Liu, and Qing Wang. Measuring memory and generalization as separable geometric channels: The topo^2 framework. *arXiv preprint arXiv:2608.30487*, 2026.

AI USE STATEMENT

Large language models were used as coding and writing assistants: generating and refactoring experiment/analysis/figure scripts, drafting and copy-editing prose, and cross-checking reported numbers against the raw result files. All experimental design decisions, judgments, and interpretations are the authors'; every number in the paper traces to a logged run, and the authors verified them against those logs. No text was included without author review.

REPRODUCIBILITY STATEMENT

The measurement protocol, including the FM0 intervention, the overlay construction and the two-point slope estimator, is specified in Sec. 3, with per-setting hyperparameters, the noise-mask reconstruction and the memory readout in App. A. The flip masks are bit-identical across seeds by construction (a seeded rebuild rather than a stored mask), so every calibration cell is reproducible from the seed alone. Table 1 reports the per-level slopes with seed counts and cross-seed CVs, and App. A states the estimator for the through-origin constant together with the anchor values used for $\mathrm{va}_{\mathrm{FM0}}$; the deprecated field that must not be read from the checkpoint metadata is flagged there as well. Deviations that a reimplementation should expect — the ±15–30% run-to-run band on absolute Betti numbers, and the pipeline-relative normalization of $H_1^{\mathrm{within}}/H_1^{\mathrm{cross}}$ — are stated in Sec. 2.2 and Sec. 3. The CIFAR-10N labels are the public release of Wei et al. (2022); the asymmetric-noise pairing is given explicitly in Sec. 4.3. Figures are produced by scripts that read the logged result files and assert the plotted values against them, so the numbers in the figures and in the text cannot drift apart.

A PROTOCOL DETAILS

CIFAR-10 / SVHN: ResNet-18 and VGG, 200–300 epochs, standard schedules, flip mask regenerated deterministically from the seed (bit-identical noise across runs). Overlay protocol: freeze FM0, unfreeze K flipped samples with clean samples present; K=half denotes the midpoint of the flipped set (flip count /2). Memory readout: $\mathrm{mem}_{\mathrm{noisy}} = \frac{1}{N_{\mathrm{noisy}}} \sum_{i:\ \mathrm{noisy}} \mathbb{1}[\hat{y}_i = \tilde{y}_i]$ (prediction matches the noisy label). All calibration numbers in Table 1 are two-point FM0→half slopes unless noted; the CIFAR-10 η=40 entry uses a 3-point line fit (K0 + Khalf + KALL).

Table 1: **Calibration of the slope k at each noise level.** Slopes are two-point FM0→half-memory unless noted (consistent with the full curve, <1%); C is the through-origin fit of k vs. $\eta/(1-\eta)$ over the uniform-flip grid. The last two rows are single-η noise-structure prediction tests of the calibrated constant (real human and confusion-pair flips; §4.3), not calibration cells.

Setting	η (%)	k	n seeds	CV / note
CIFAR-10	20	0.110	3	CV 12.6%; out-of-grid, +14% off law (§5.4)
CIFAR-10	30	0.181	3	3.5%
CIFAR-10	40	0.258	3	2.2%; 3-pt fit
CIFAR-10	50	0.410	3	5.7%; 8-pt curve fit 0.386 (s42)
CIFAR-10	60	0.559	3	2.7%
SVHN	20	0.092	3	CV 5.4%; out-of-grid, ratio 0.93; 5-pt C=0.397
SVHN	30	0.166	3	5.0%
SVHN	40	0.253	3	4.4%
SVHN	50	0.400	3	3.5%
SVHN	60	0.600	3	4.5%
CIFAR-100	30	0.206	3	2.9%; Sec. 4.3.b
CIFAR-100	50	0.366	3	framework record; 3-seed mean 0.371 (Sec. 4.3.b)
VGG(SVHN)	30	0.152	3	6.9%; Sec. 4.3.c
VGG(SVHN)	50	0.376	3	3.9%; Sec. 4.3.c
VGG(SVHN)	60	0.544	3	3.7%; Sec. 4.3.c
noise-structure anchors at $\eta \approx 40\%$ (prediction tests of the calibrated constant; not in the C fit; §4.3)				
CIFAR-10N (real human)	40.2	0.271	3	ratio 1.043; on-law
CIFAR-10 asym (confusable pairs)	40	0.340	3	ratio 1.320 (+32%)

B REPRODUCIBILITY NOTES

- FM0 anchor accuracies used as va_{FM0}: CIFAR-10 η=50 0.9223; SVHN η=50 0.9629. For the SVHN/VGG FM0 anchors, $\text{mem}_{\text{noisy}}$ was not logged at train time and the nominal FM0 value 0.004 is used (an earlier logged memory field was affected by a readout bug and is discarded throughout; cf. the selective-forgetting entry of Zhang et al. (2026)'s falsification ledger).
- The calibration constant is estimated as $C = \frac{\sum_i x_i k_i}{\sum_i x_i^2}$, $x_i = \eta_i/(1-\eta_i)$, over the in-grid settings of each dataset ($\eta \in \{30, 40, 50, 60\}$).

C FALSIFIED ALTERNATIVES AND NON-CLAIMS

- **FM0 is not SOTA**; FM0-hat is mid-tier on CIFAR-10 η=40 (Sec. 6).
- H_1^{within} **is not invariant** at value 30; it is zero-deviant from the random baseline (Zhang et al., 2026).
- **Selective forgetting / ring overload are falsified** (the falsification ledger of Zhang et al. (2026)).
- **Cross-H_1 is not a deployed generalization predictor** (Sec. 7).
- The **concentration mechanism** ("label spread *across distinct classes* matters") was falsified by CIFAR-100; the surviving mechanism is the sample concentration ratio (Sec. 5.5). Flip structure is a separate, weak axis (Sec. 4.3).
- C **as a capacity-independent constant is retracted**: C is a capacity-dependent effective coefficient; the four-setting agreement at 0.37–0.38 reflects the shared ~11M reference capacity (Zhang et al., 2026).

D SUPPORTING FIGURES

The three figures below support Sec. 4.3 (functional form), Sec. 5.4 (low-noise residuals) and Sec. 5.6 (capacity dependence); they are placed here for the main-text page budget.

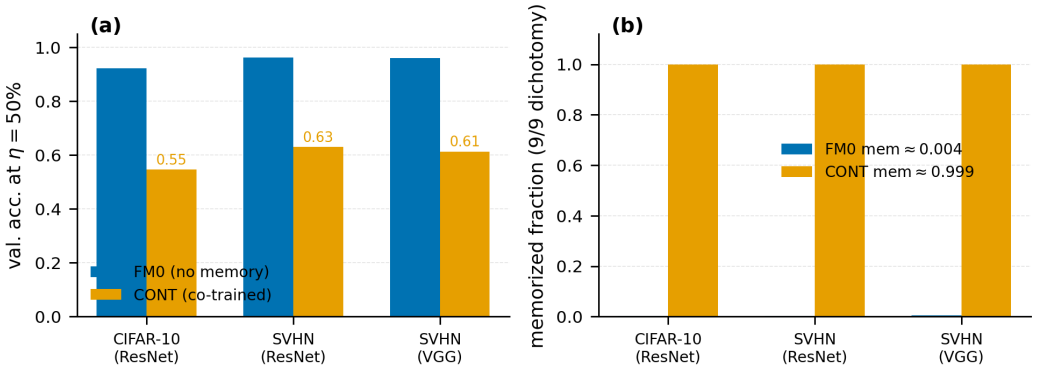


Figure 4: **FM0 reaches each setting's ceiling; the memory readout dichotomizes.** (a) validation accuracy at η=50: FM0 (no memory) attains each setting's ceiling; CONT (co-trained) pays the memorization cost. (b) The dual-scope readout: FM0 mem ≈ 0.004 vs. CONT mem ≈ 0.999, 9/9 setting-seed dichotomy.

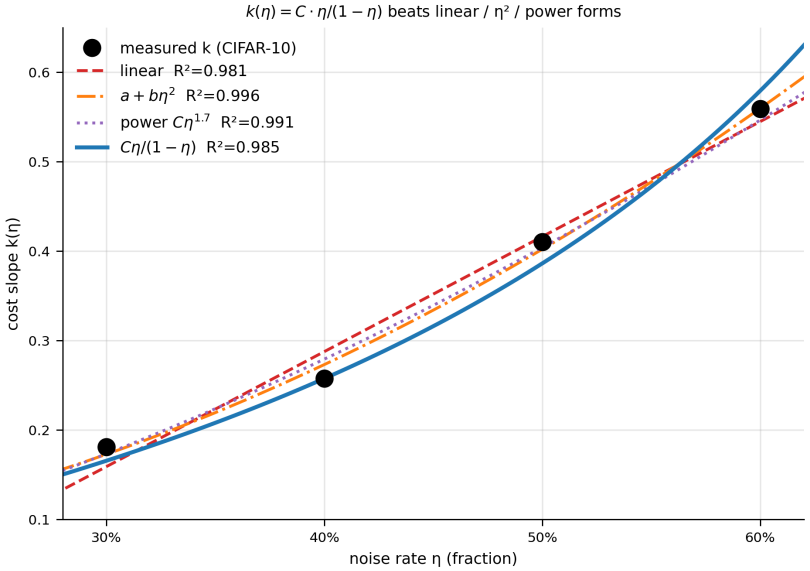


Figure 5: **The $\eta/(1-\eta)$ form versus alternatives** on CIFAR-10 ($\eta \in \{30, 40, 50, 60\}$): linear R^2=0.981, power-law 0.991, η^2 0.996, $\eta/(1-\eta)$ 0.985 — a statistical tie with η^2, retained on mechanistic grounds (the noise:clean ratio of Sec. 5).

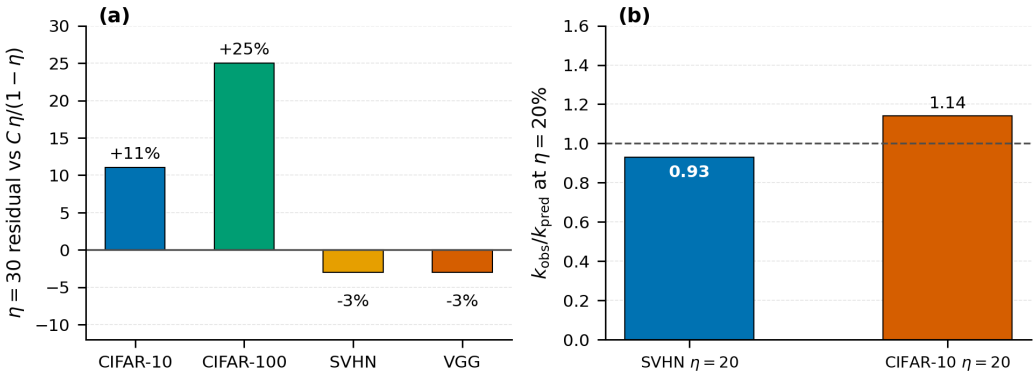


Figure 6: **Honest boundaries of the law.** (a) The η=30 single-point residual against $C \cdot \eta/(1-\eta)$ per setting: CIFAR-10 +11%, CIFAR-100 +25%, SVHN −3%, VGG −3% — the $\eta/(1-\eta)$ form is first-order. (b) η=20 extrapolation (3-seed mean ± sd): SVHN ratio 0.93, CIFAR-10 ratio 1.14 — both near the law, with the low-noise residual scatter larger than in-grid.

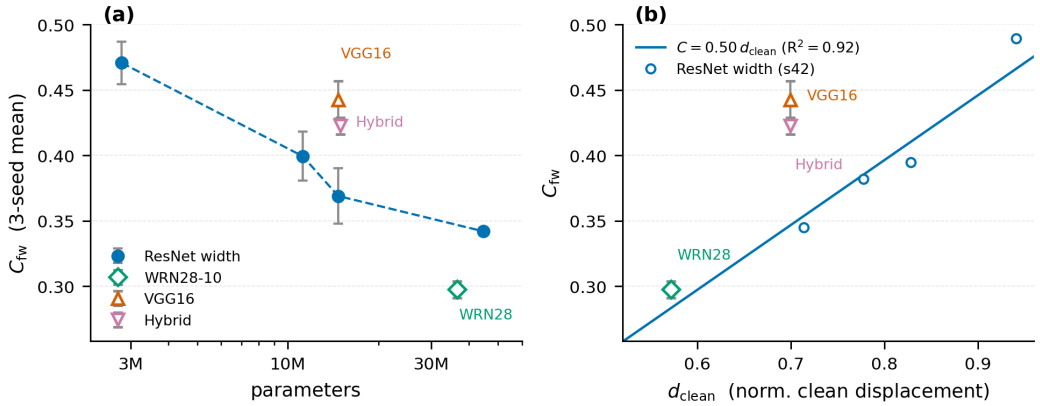


Figure 7: **Capacity dependence of C and its d_{clean} reduction.** (a) C_{fw} vs. parameters (3-seed mean ± sd): the ResNet18 width family falls monotonically 0.471 → 0.342 ($w0.5$ → $w2.0$, dashed guide); WRN28-10 (36.5M) sits below the trend, VGG16 and the attention-hybrid above it. (b) C vs. d_{clean}. The width family reproduces the through-origin regression of Zhang et al. (2026), C=0.4955 d_{clean} (R^2=0.92; family points shown at that paper's $s42$ reference). The 3-seed cross-family anchors land on the line for WRN28-10 (±5%, absorbed) but +28%/+22% above it for VGG16 and the hybrid — d_{clean} is a verified mediator within the residual-conv family.