IDENTITY SUPERVISION FOR READABLE INTERMEDIATE STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

How do looped transformers organize intermediate states that new downstream computations can use across source views? We investigate this question through controlled studies of modular addition and two-hop relational composition, comparing weight-tied and untied transformers at matched effective depth. Our analysis combines temporary supervision interventions, state diagnostics across checkpoints, and tests of new readers and retained downstream computation. These tests distinguish identity information readable from a state from its usefulness to a specified attribute reader, and examine the roles of supervision location, gradient exposure and source variation. The tested identities are known, and the later reader generalizes their observed labels across source views. In relational composition, direct identity supervision yields gains of 5.2–7.6 percentage points across seven settings under a common bounded gradient rule, while several comparisons with fixed calibration remain unresolved. In modular addition, near-perfect affine identity reading coexists with substantially weaker random-attribute reading. Depth-by-strength comparisons qualify a pure supervision-distance explanation, and matching identity grouping reveals different costs for source-factor reading. Together, these results characterize concrete dependencies and limits of intermediate-state use rather than a universal benefit of weight sharing. The study connects source-view generalization to analysis of internal states while distinguishing reader-specific accessibility from stronger claims about computational mechanisms.

1 INTRODUCTION

A hidden state can support its model's original prediction while remaining difficult for another module to read. Learned connectors link frozen encoders to new consumers, as in BLIP-2 (Li et al., 2023); early-exit networks attach classifiers before the original output (Kaya et al., 2019). Both motivate a concrete design choice: which state should be exposed, and which supervision will make that state useful? A decoder may finish an identity computation that a newly attached reader must otherwise learn again.

We study this question in a controlled, known-identity setting. A source x expresses an intermediate identity m; different sources can express the same identity. After training relational composition or decomposed modular addition, we freeze an intermediate state and fit readers for new attributes of those identities. Every tested identity has an observed attribute label. The challenge is to transfer that label from a few source realizations to other realizations, which we call *source-view generalization*. Original atomic training includes these held sources. This setting isolates accessibility across views; it does not reproduce the unknown inputs or scale of pretrained connectors.

Our first question concerns *identity access*. We compare identity cross-entropy through the original decoder with a temporary affine head at the exposed state, Direct-ID. Deep supervision already provides this mechanism, including classifiers discarded after training (Lee et al., 2015). The contribution here is the controlled comparison of a later task whose labels never supervise the source model: nonlinear attribute reading, affine identity reading and affine attribute reading need not agree.

Our second question concerns *gradient exposure*. Moving the same fresh head downstream changes its prediction residual, learned weights and backward transformation. Matching gradient norms at the intervention's start does not keep these quantities matched during training. We therefore cross

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.

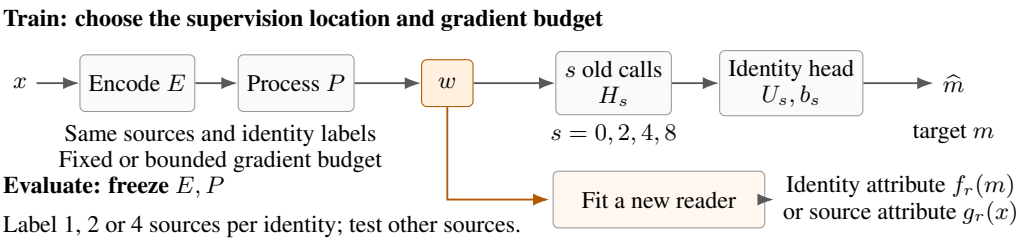

Figure 1: **Supervision depth and readout are separate choices.** A temporary identity head supervises the exposed state through s original consumer calls. At $s = 0$, it is Direct-ID. Evaluation always starts from the same boundary w and uses new labels. Pair contraction and SupCon provide complementary objectives directly at w.

depth with supervision strength, repeat the scan with distinct parameters, and compare a bounded per-update controller for all competing losses. These controls ask how much of an apparent distance effect survives changes in the pressure applied to the interface.

Our third question concerns *compatibility*. Greater class separation can help a newly fitted classifier recover an original classifier's performance (Kornblith et al., 2021), yet useful concentration depends on the target task (Galanti et al., 2022; Hui et al., 2022). We compare readers requesting identity attributes with readers requesting source attributes that may differ within one identity. Matching training-state grouping before evaluating either target tests whether one scalar grouping score captures the resulting tradeoff.

The resulting account is conditional and operational: specify the boundary, the auxiliary gradient budget and the distinctions the next reader needs. The original decoder remains a reference when available. The experiments quantify the consequences of exposing another state; they do not require replacing a decoder that already solves its task.

2 Related work

Supervised interfaces. Deep supervision trains intermediate classifiers (Lee et al., 2015), and multi-exit networks use them to terminate inference early (Kaya et al., 2019). Direct-ID uses this established mechanism. We discard the identity classifier and fit a new attribute reader after freezing the source model; its attribute labels never supervise source training. We study how supervision location changes this later interface and the source distinctions it preserves. Connectors between frozen components (Li et al., 2023) and model stitching (Bansal et al., 2021) motivate interface reuse; linear probes measure classifier-specific accessibility (Alain & Bengio, 2016). Gradient balancing is also established (Chen et al., 2018). Our bounded per-update control targets a common fixed ratio of gradient norms and caps the loss coefficient, whereas GradNorm adapts targets to relative task-training rates.

Class separation and transfer. In Section 3.4 and Figure 6 of Kornblith et al. (2021), greater class separation narrows the gap between an original ImageNet classifier and a linear classifier relearned on the same classes; stronger separation can nevertheless impair transfer to different tasks. Neural-collapse analyses connect concentration to few-shot transfer under distributional assumptions (Galanti et al., 2022), while Hui et al. (2022) distinguish training collapse from test behavior and show cases where further collapse hurts transfer. The tunnel effect describes later layers compressing discriminative features while degrading out-of-distribution performance (Masarczyk et al., 2023). These findings motivate measuring geometry and readout together. Our retained-decoder comparison measures recovery along a fixed computation.

Which distinctions to preserve. Contrastive objectives can retain useful information beyond source labels (Islam et al., 2021), but can also suffer dimensional collapse (Jing et al., 2022). Within-class grouping and global loss of feature dimensions are different properties. Useful invariances depend on the downstream task (Tian et al., 2020; Speicher et al., 2024). Augmentation-aware su-

pervision preserves transformation details (Lee et al., 2021); equivariant contrastive learning combines sensitivity to selected transformations with invariance to others (Dangovski et al., 2022). Our identity-dependent and source-dependent readers test whether similar identity grouping entails similar costs for other distinctions on the same source population.

Controlled reasoning settings. Compositional benchmarks separate learning facts or rules from combining them (Lake & Baroni, 2018; Keysers et al., 2020; Press et al., 2023). Relational grokking exposes intermediate fact use (Wang et al., 2024); modular arithmetic supports analyses of learned algorithms (Power et al., 2022; Nanda et al., 2023; Zhong et al., 2023). Recurrent-depth models make repeated computation explicit (Dehghani et al., 2019; Saunshi et al., 2025). We use these settings to control identities and source variation. Random-attribute evaluation measures generalization across sources for observed identity–reader labels, separately from the original tasks' compositional generalization.

3 A CONTROLLED INTERFACE AND ITS SUPERVISION

We distinguish the original composition task from a later source-view evaluation. Original training teaches intermediate identities and their composition; the later evaluation freezes the source computation and asks how a new reader generalizes observed labels to other sources of the same identity. Appendix A gives the full specification.

3.1 SOURCES, IDENTITIES AND COMPUTATION

Relational composition. We adapt the two-hop setting of Wang et al. (2024), also studied with recurrent depth and targeted composition supervision (Kohli et al., 2026; He et al., 2026). A source fact $T(h, r_1) = m$ supplies an identity, and a reader asks for $T(m, r)$. Each of 500 entities has ten distinct outgoing labels from 50 relations and independently sampled targets: 5,000 facts and 50,000 ordered paths. All facts receive atomic supervision. We designate 250 facts as absent from compositional training, and split paths with two ordinary facts 95:5 into training and an in-role test. The role-held-out test uses a designated first fact and an ordinary second fact; paths with designated second facts are excluded. This tests a new compositional role for a known fact, not inference of an unknown random edge.

Decomposed modular addition. Modular arithmetic supports controlled studies of grokking and learned algorithms (Power et al., 2022; Liu et al., 2022; Nanda et al., 2023; Zhong et al., 2023). We factor addition into a source sum and a subsequent reader addition. For each $m \in \mathbb{Z}_{257}$, choose eight distinct prefixes (a, c) with $a + c = m$, split into four A and four B sources. Atomic training teaches all 2,056 prefixes to predict m. Split ordered (m, b) pairs into 33,024 training and 33,025 held-out pairs before expanding sources. Composition training uses A sources, giving 132,096 queries $[a, c], b \mapsto m + b$; evaluation uses B sources with held-out canonical pairs, giving 132,100 queries. Arithmetic is modulo 257, and commuted pairs may fall in different splits.

The boundary being exposed. Following recurrent-depth models (Dehghani et al., 2019; Geiping et al., 2025), the default model reuses one causal pre-RMSNorm Transformer block:

$$z = E_\theta(x), \qquad w = P_\theta(z), \qquad \ell = C_\theta(w, r). \tag{1}$$

Encoding calls the block twice on the two-token source and retains its final position. Processing makes two calls on that position alone. Consumption appends the reader and makes eight further calls. Atomic queries omit the reader but keep the twelve-call schedule. With one token, processing attention has weight one: this stage is a residual MLP with a normalized linear attention path. Widths 128/256 give 331,904/919,808 parameters on graphs/addition. Both models use four attention heads, GELU width $4d$, learned token and position embeddings, final RMSNorm and a bias-free output head. There is no dropout or input reinjection. A second architecture assigns a distinct block to each of the twelve calls, giving an untied twelve-block model with 2.50M/9.58M parameters (Appendix A.1).

3.2 MOVING IDENTITY SUPERVISION ALONG THE COMPUTATION

At update 5k, copy a native model and its AdamW state into each comparison branch. Add an auxiliary loss for 5k updates, then continue native cross-entropy to 200k. This short early intervention asks whether changing an interface during learning affects its later use. Each auxiliary minibatch contains the same 64 legal same-identity source pairs, or 128 endpoint occurrences, using existing atomic-training labels. Designated graph sources and B addition sources never enter these pairs.

Atomic-CE predicts identity through the original eight-call atomic consumer, final normalization and trained output head. **Direct-ID** predicts it with a temporary affine head directly on w. To separate these endpoints systematically, let H_s apply the first s original atomic-consumer calls, with $H_0(w) = w$, and train the same fresh-head family at every depth:

$$\mathcal{L}_{\mathrm{ID},s} = \mathbb{E}_{x,m}\,\mathrm{CE}\big(U_s H_s(w) + b_s, m\big), \qquad s \in \{0, 2, 4, 8\}. \tag{2}$$

Depth zero is Direct-ID. Depth eight still uses a new head, so it remains distinct from Atomic-CE's original head. The auxiliary gradient passes through detached consumer parameters into P; encoding is detached. The head learns with its own optimizer and is discarded before evaluation. All depth comparisons subsequently read the original boundary w, not $H_s(w)$. Appendix A.3 specifies this separation.

Contraction groups equivalent sources without predicting their identity. For $\delta_{AB} = \|\,\mathrm{sg}(z_A - z_B)\|_2$, its pair loss is

$$\mathcal{L}_A = \mathbb{E}_{A,B}\frac{[\|P_\theta(\mathrm{sg}\, z_A) - P_\theta(\mathrm{sg}\, z_B)\|_2 - \rho\delta_{AB}]_+^2}{\delta_{AB}^2 + \epsilon}, \tag{3}$$

where ϵ stabilizes the denominator. It penalizes sampled pair distances above a fraction ρ of their input distance. **SupCon** instead applies supervised contrastive loss to unit-normalized endpoint states (Khosla et al., 2020), using all same-identity nonself positives, all other occurrences in the denominator, and no projection head. The defaults are $\rho = 0.5$ and temperature $\tau = 0.1$.

Fixed and bounded gradient budgets. Native CE differentiates the complete model. Let $g_{N,t}$ be its full source-model gradient and $g_{A,t}$ the auxiliary processing-route gradient, padded with zeros for other parameters. The fixed baseline uses $\lambda_0 = \gamma\|g_{N,0}\|/\|g_{A,0}\|$, calibrated on reserved minibatches at the parent checkpoint. The bounded controller recomputes

$$\lambda_t = \min\left\{c_{\max}\lambda_0,\ \frac{\gamma\|g_{N,t}\|}{\|g_{A,t}\| + \epsilon_g\|g_{A,0}\|}\right\}, \tag{4}$$

with $c_{\max} = 10$ and $\epsilon_g = 10^{-6}$. Coefficients and norms are detached; the source optimizer receives $g_{N,t} + \lambda_t g_{A,t}$, globally clipped to norm one, with its inherited AdamW state. The new head has a separate optimizer and unscaled CE. This targets an auxiliary norm ratio of at most γ, without claiming equal update directions or optimizer displacement. The cap limits amplification as auxiliary gradients vanish. We apply the same rule to Atomic-CE, Direct-ID and Contraction; default $\gamma = 0.1$. Appendix A.3 specifies zero-gradient cases and diagnostics.

Shared parameters still couple the forward computations of E, P and C even though the auxiliary derivative follows only the P calls. We therefore cross $s \in \{0, 2, 4, 8\}$ with fixed $\gamma \in \{0.1, 0.5\}$ and bounded $\gamma = 0.1$ in both shared and fully untied models. Native training continues to update every stage in both architectures.

Transfer-level sweeps vary Direct-ID and Contraction over $\gamma \in \{0.05, 0.1, 0.2, 0.5\}$, Contraction over $\rho \in \{0.5, 0.7, 0.9\}$, and SupCon over $\tau \in \{0.05, 0.1, 0.2\}$. Functional consistency, random-direction contraction, mismatched pairs and spectral regularization provide complementary controls. Appendix A.2 defines them and quantifies their different pair constraints. Five-thousand-update interventions starting at 40k or 160k, and continuous auxiliary training from 5k, test dependence on the early window (Appendix A.3).

3.3 GENERALIZING LABELS ACROSS SOURCE VIEWS

At 200k freeze embeddings, E and P. Four independently permuted balanced maps $f_r(m)$ assign each identity to one of 16 classes. Every test identity–reader pair has a training label. We vary

only the sources carrying those labels: one, two or four nested views per identity. For graphs, retain identities with at least five ordinary and one designated source, reserve one ordinary source, select four for training, and test all designated sources. This retains 174–198 identities per task. Addition trains on A sources and tests all B sources. No test prefix enters downstream fitting. The task is source-view generalization of known attributes, rather than unseen-identity or new-rule learning.

A **fresh consumer** uses a new eight-call block, four reader embeddings and a 16-way head, trained for 2,000 updates. A **logistic attribute probe** fits four independent affine multinomial classifiers. With four views, four-fold validation holds out one training source per identity, selects the penalty by validation CE, and refits all four views. A **ridge identity probe** predicts the original identity using 500 graph or 257 addition outputs; it selects its penalty by validation one-hot squared error under the same source folds. Its graph evaluation contains only the 174–198 eligible identities. A **nearest-view reader** retrieves the observed label of the nearest training state. Each fit uses a scalar training RMS; no test statistic selects a parameter.

These are different learning problems: affine identity discrimination does not make every random 16-class partition affine. All identity probes are newly fitted at 200k, after the temporary training heads have been discarded. Appendix A.4 gives objectives, grids and fitting diagnostics. Fixed-penalty attribute ridge probes remain useful for the compatibility comparison below and for full-budget sensitivity checks. One view cannot supply a held-training-source validation fold. Additional fitting at matched epoch counts separates source coverage from reader training duration (Appendix A.4).

3.4 Separating grouping, retained computation and task compatibility

We also fit probes after $k \in \{0, 2, 4, 8\}$ frozen original consumer calls. Here k changes the *readout*, whereas s in Equation 2 changes *training supervision*. The complete original decoder plus observed-label lookup is a separate reference. Because the tested identities and their new labels are known, successful identity decoding suffices for that reader.

For a conflicting task, four new maps label the first graph relation r_1 or addition's first operand a. Such labels can differ between sources of one identity. We fix evaluation to factor categories observed in one-view training and compare identity attributes and source attributes on exactly the same test sources, using four training views and the same fixed-penalty attribute ridge rule ($\alpha = 0.01$) (Appendix A.5).

Geometry is measured by within-identity scatter W, between-centroid scatter B, their ratio W/B, and identity-neighbor retrieval. We relate these diagnostics to fresh-consumer performance across all nine objectives, treating the objectives as a fixed comparison set and tasks as the replication unit (Appendix A.6). Finally, we match a Contraction setting to Direct-ID using only four-view training geometry before examining held-source outcomes (Appendix A.6). This tests whether one grouping score summarizes the compatibility tradeoff; it does not equate all geometric properties of two representations.

What changing supervision depth changes. For a minibatch of n sources, the processing-route gradient is

$$g_{A,s} = \frac{1}{n} \sum_{i=1}^{n} J_{P,i}^{\top} J_{H_s,i}^{\top} U_s^{\top} r_{s,i}, \qquad r_{s,i} = p_{s,i} - e_{m_i}. \tag{5}$$

Here $r_{s,i}$ is the prediction residual, U_s the current head, and the Jacobians differentiate the specified auxiliary call path. Depth can change all three factors after J_P, as well as cancellation between examples. Residual blocks need not attenuate gradients, and a learned head can compensate for an invertible consumer transformation. Equal initial norms therefore do not isolate supervision distance. Our bounded control checks sensitivity to gradient magnitude; even equal norms leave direction and optimizer dynamics unmatched. Appendix C, Equation 16, gives a local example with identical losses and gradient norms but different coordinate pressure, together with a compensating-head case.

Grouping answers a different question. Two banks can have the same W/B while retaining different source factors (Appendix C). We consequently measure identity readability and source-factor compatibility separately.

4 RESULTS

The fixed-coefficient baseline uses sixteen tasks per family and four source initializations; each fresh reader additionally averages three initializations. Its four one-view comparisons of Direct-ID or Contraction against Atomic-CE use Holm correction. The common readout panel uses the first eight tasks and two source initializations. Controller and depth comparisons use eight tasks and four source initializations. Differences are paired within task before forming 95% intervals; follow-up intervals are descriptive. Appendix B.1 specifies aggregation, and Appendix B gives complete results.

4.1 IDENTITY ACCESS AND ATTRIBUTE READING ARE DIFFERENT PROPERTIES

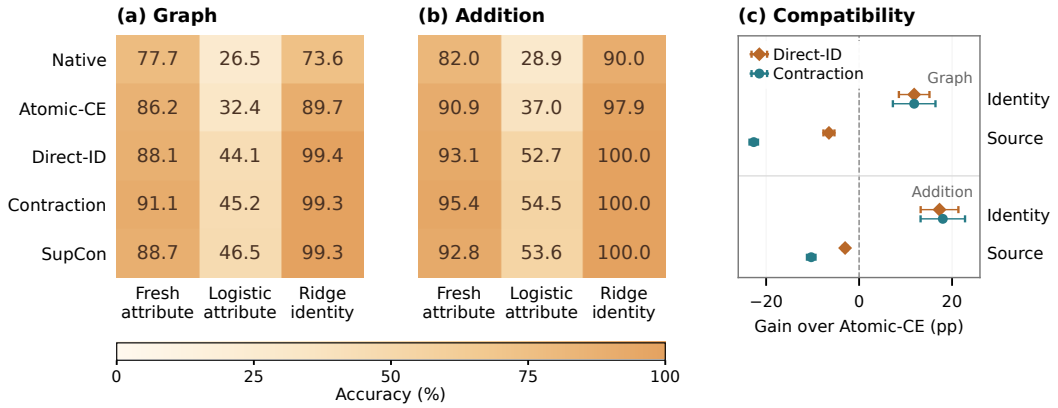

Figure 2: **What a new reader can recover from the same state.** (a,b) Four-view accuracy (%) on a shared color scale in fixed-coefficient banks: a fresh nonlinear attribute reader, a source-CV logistic attribute probe and a source-CV ridge identity probe. The last column predicts identity, not the random attribute. (c) Paired changes from Atomic-CE for identity-attribute and source-factor reading on the factor-covered subset, using fixed ridge $\alpha = 0.01$. All panels use eight tasks and two source initializations; intervals are paired 95% task intervals.

The baseline one-view comparison gives Direct-ID gains of +5.34 $[3.08, 7.60]$ points on graphs and +7.64 $[5.11, 10.18]$ on addition over Atomic-CE. Contraction gives +8.06 $[5.58, 10.54]$ and +8.87 $[5.35, 12.38]$. The new attribute labels never supervise the source model. Thus a temporary identity objective can benefit a later reader even though its own head is discarded. Figure 2 compares five objectives on the same smaller cohort; Table 3 includes one-view accuracy and the original composition task, while Table 4 retains all nine objectives and budgets.

A direct test of affine identity readability. At 200k, newly fitted four-view identity probes reach 99.35% and 99.35% on Direct-ID and Contraction graph states, compared with 89.68% on Atomic-CE. The corresponding addition accuracies are 99.99%, 100.00% and 97.90%. Direct-ID's paired identity-probe gains are +9.67 $[3.71, 15.62]$ and +2.10 $[-0.31, 4.50]$ points: the graph difference is resolved, while addition is close to ceiling. These probes test the final state's accessibility, not retention of the particular head discarded at 10k.

High identity accuracy does not imply equally high attribute accuracy. The four-view logistic attribute probes in Figure 2 remain substantially below the identity probes; arbitrary partitions of identities need not be affine (Appendix C). Both interface objectives improve the logistic result over Atomic-CE, with no reliance on the poorly regularized one-view regime. Every identity CV fit selects the largest penalty in the declared grid, $\alpha = 1$; this reports a finite-grid result, not globally optimal probe performance. Appendix A.4 gives training fit and the full identity grid.

Why extra atomic training is a substantial intervention. Native training already includes all atomic facts, but samples uniformly from the union of atomic and composition queries. A full 512-query batch contains only 53.5 expected atomic queries on graphs and 7.85 on addition. Each pulse

step adds 128 atomically labeled endpoint occurrences, about 2.4 and 16.3 times those expected counts. These are exposure counts, not gradient weights. The original decoder's atomic accuracy at the 5k parent is 70.05% on graphs and 75.77% on addition. Composition-training accuracy is about 37% in both families; the in-role tests are about 5% (Appendix B.3). The intervention therefore begins before either task is solved.

Targeted reweighting and inherited optimizer state offer a plausible route for the large Native–Atomic gap. Parent accuracy alone cannot determine later auxiliary CE or its gradient: the pair sampler reweights identities, and accuracy does not measure confidence. The pulse diagnostics below address that distinction directly. The fixed-bank temporal analysis in Appendix B.3 also separates endpoint readability from the geometry along a specified path between checkpoints.

4.2 A COMMON GRADIENT BUDGET IMPROVES ROBUSTNESS AND QUALIFIES THE DEPTH EFFECT

Table 1: Direct-ID minus Atomic-CE on one-view fresh-consumer accuracy, in percentage points with paired 95% task intervals. Each comparison uses the same coefficient rule for both objectives: fixed initial calibration or the bounded adaptive rule. Eight tasks, four source and three reader initializations.

Setting	Graph		Addition	
	Fixed	Bounded	Fixed	Bounded
Shared, 5k–10k	4.5 [1.2, 7.7]	6.1 [4.7, 7.5]	8.7 [3.8, 13.5]	6.3 [4.7, 8.0]
40k–45k	2.8 [-1.7, 7.4]	6.7 [4.6, 8.7]	6.4 [0.1, 12.7]	7.1 [5.6, 8.7]
160k–165k	3.1 [-0.7, 6.9]	5.9 [3.5, 8.2]	7.6 [1.2, 14.1]	7.6 [6.3, 8.8]
5k–200k	4.7 [0.7, 8.8]	7.6 [6.1, 9.0]	9.3 [2.7, 16.0]	7.9 [4.8, 10.9]
Three stages	4.4 [-2.1, 10.8]	6.7 [3.6, 9.9]	6.7 [2.9, 10.6]	9.6 [6.8, 12.5]
Other decay	1.8 [-2.9, 6.4]	5.2 [3.2, 7.2]	7.1 [1.3, 12.9]	7.4 [5.1, 9.8]
Twelve blocks	1.2 [-3.4, 5.8]	5.4 [3.8, 7.0]	8.0 [2.7, 13.4]	8.3 [5.7, 10.9]

A coefficient calibrated once equalizes only the initial norm ratio. If the native and auxiliary gradients evolve differently, nominally equal γ values can deliver different cumulative pressure. Table 1 compares fixed calibration with the bounded controller of Equation 4, always subtracting Atomic-CE under the *same* control rule. Fixed Direct-ID has positive graph means throughout, but several intervals span zero. With bounded control, its graph gains are more consistent across the early, middle and late windows, continuous training, distinct-stage and untied architectures, and alternative decay. The paired change in the Direct-ID–Atomic contrast is +1.62 $[-1.22, 4.46]$ points for the early shared model and +4.20 $[0.26, 8.14]$ for the untied model. Only the untied graph interaction excludes zero; the other graph interactions and all addition interactions remain unresolved. Thus consistent positive gains under one rule do not imply that it significantly improves every regime.

The interpretation is specific: gradient-budget choice can matter enough to alter apparent regime sensitivity. Eight tasks do not separate every smaller architecture or schedule interaction. Continuous supervision also uses 195k auxiliary updates rather than 5k, while later windows leave less time before evaluation. The controller is a training rule shared by all objectives; it does not establish free gains, equal optimizer displacement or unchanged original-task accuracy. Appendix B.3 reports Contraction under the same control rule and the complete depth-by-strength design.

Depth crossed with dose and parameter sharing. Using the same new head family removes the original-versus-new-head confound. Under fixed $\gamma = 0.1$, moving from $s = 0$ to $s = 8$ costs +5.35 $[3.93, 6.76]$ graph points in the shared model. Increasing the dose at deep exits recovers part of this difference. Under bounded control the corresponding separation is +1.91 $[0.57, 3.25]$ points; in the untied model it is +1.06 $[0.14, 1.98]$. Addition also retains depth-zero advantages; the reduction of its untied depth gap remains uncertain. Table 9 reports every cell. The untied scan eliminates auxiliary parameter sharing across stages, although native training still updates every stage.

This changes the mechanistic conclusion. Supervision distance is one part of a coupled intervention that also changes residual error, head adaptation and gradient transport. The pulse CE and raw-gradient profiles show why a one-time normalization cannot hold pressure fixed (Figure 4). Larger

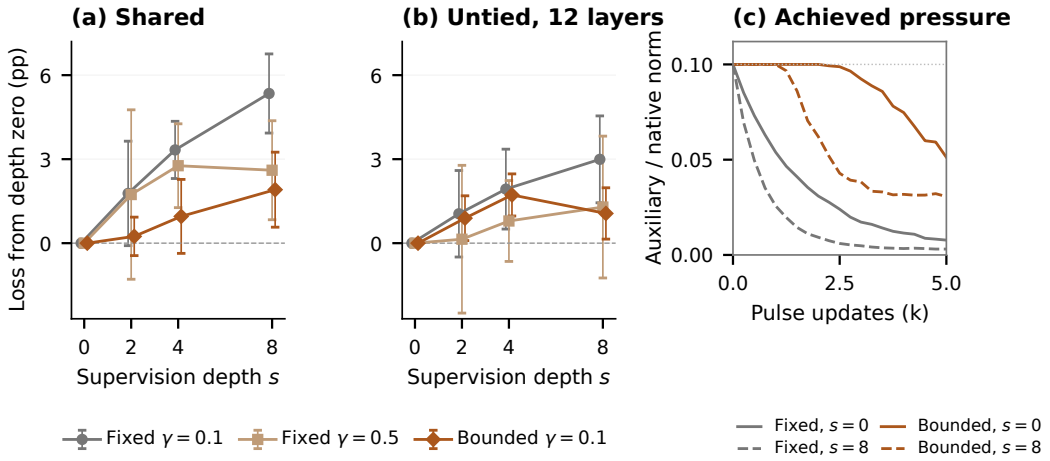

Figure 3: **Supervision depth interacts with gradient strength.** (a,b) Graph fresh-reader loss relative to the same control's $s = 0$ head, for shared and twelve-layer untied models. All readers use w; intervals are paired across eight tasks. (c) Mean achieved auxiliary/native gradient-norm ratio during the shared graph pulse, for depths zero and eight. Fixed calibration can decay away from its initial target; bounded control compensates subject to its coefficient cap. Appendix B.3 gives both task families, raw losses and norms, and cap diagnostics.

deep-head doses and bounded control shrink the depth contrast, so the original five-point separation should not be attributed wholly to distance. The cap binds for 32.7% of updates at graph $s = 0$ and 70.3% at $s = 8$, leaving mean achieved ratios of 0.088 and 0.059. Remaining differences can therefore still involve magnitude, as well as direction or optimizer history; they do not identify Jacobian attenuation. The local calculation in Appendix C illustrates that distinction without assuming every residual block contracts gradients.

4.3 Similar grouping can preserve different useful distinctions

Across the nine fixed-coefficient objectives, better identity grouping strongly predicts fresh-consumer accuracy. The task-aggregated correlations with $\log(W/B)$ are -0.91 $[-0.93, -0.88]$ on graphs and -0.89 $[-0.91, -0.87]$ on addition. Figure 6 shows the task means and objective averages. This agrees with established class-separation results (Kornblith et al., 2021; Galanti et al., 2022). Yet the association is incomplete: a calibration fitted to the other eight objectives overpredicts SupCon by about six and seven points. Appendix B.4 reports this diagnostic, and Equation 17 in Appendix C explains why nearest-view random-attribute accuracy largely follows identity retrieval.

Compatible and source-dependent readers. Figure 2(c) compares identity attributes with source attributes on identical factor-covered test sources. Contraction improves identity reading but can remove distinctions useful for the source-factor reader. Direct-ID has smaller source-factor costs while retaining identity-attribute gains. These are objective differences within a reader; the two label families need not have equal absolute difficulty. The evidence concerns fixed-coefficient banks and is separate from the adaptive-controller endpoints above.

Simply weakening contraction might recover the same tradeoff. We therefore select a Contraction setting from the declared ρ and γ grid to match default Direct-ID's four-view *training* W/B, before examining held-source labels. Direct-ID exceeds matched Contraction on source-factor reading by +10.91 $[4.76, 17.05]$ points on graphs and +5.36 $[2.52, 8.21]$ on addition. Its identity-attribute differences are +2.15 $[0.92, 3.38]$ and +1.50 $[-2.20, 5.19]$. Both graph intervals favor Direct-ID: improved factor reading accompanies improved compatible reading. Figure 7 displays every grid point and Table 12 gives paired intervals. Matching is approximate and scalar; it equates neither all geometry nor class margins.

Coverage and retained computation delimit the benefit. More source views improve absolute reading, while narrowing several objective gaps. To distinguish coverage from fitting time, we extend four-view fresh-reader training to the approximate epoch count of the one-view condition. Contraction's four-view gains then remain +4.78 $[0.40, 9.15]$ and +2.64 $[-1.24, 6.52]$ points; the mean narrowing survives, though its graph interval spans zero (Appendix B.3). Fresh-consumer training accuracies remain above 99%, which alone does not settle later generalization.

The original decoder plus observed-label lookup is nearly perfect for identity attributes. That is expected for atomically trained sources with known labels. Keeping only a prefix of that decoder gives intermediate readout choices (Table 6); this changes evaluation depth, separately from training supervision depth. The decoder is therefore a strong compatible-task reference, while its identity output alone cannot distinguish sources sharing that identity (Appendix B.2).

The intervention's cost also depends on location: increasing s from zero to eight raises dense forward MACs per auxiliary endpoint from 1.25M to 2.82M on graphs and 4.79M to 11.09M on addition. These counts exclude backward operations and optimization. Bounded control requires separate gradient norms and does not remove this cost. Appendix A.6 gives the accounting and reference implementation.

5 Discussion

A reusable interface depends on identity access, training exposure and the distinctions a later reader needs. Direct-ID uses an established auxiliary classifier to improve the state read by a new task. The comparisons sharpen that prescription: a head's location and its effective gradient strength interact, and a coefficient calibrated once may give uneven pressure over the intervention. A common bounded controller makes the graph comparisons more consistent, while the depth-by-strength scan weakens a purely geometric explanation of supervision distance.

These assays define use relative to a continuation. An identity probe fits a new affine map to recover m; the fresh-consumer assay learns $f_r(m)$ from selected source views. Retained-computation probes test accessibility after a fixed prefix of the original decoder. Their scores do not establish a common algorithm or identify the features on which native composition causally relies. That question requires interventions that hold the receiving context and continuation fixed while changing a proposed source contribution. A stronger functional-transfer test would hold the native continuation fixed, replace its input with a state from another source context, and evaluate without reader refitting. This would test interchangeability under that continuation; algorithm identity would require additional evidence. Tied and untied models here match the twelve-call schedule and interface, while their unequal parameter counts preclude a capacity-matched architecture ranking.

The compatibility study adds a separate design criterion. Strong identity grouping predicts relearning performance across objectives, but comparable grouping need not preserve the same source distinctions. In the fixed-coefficient banks, Direct-ID retains more source-factor readability than matched Contraction; on graphs it also gives better compatible reading. The adaptive-controller comparisons evaluate fresh-reader endpoints separately, so this compatibility advantage should not automatically be assigned to their states.

A decoded identity is sufficient for every identity-attribute table here, but cannot distinguish two sources with the same identity and different factor labels. A system serving both requests must retain those distinctions somewhere. Evaluating both readers makes that requirement visible; neither a single transfer score nor a scatter ratio captures it.

The scope remains small and controlled: known atomic sources, known downstream identity labels, two synthetic task families and single-token processing states. The untied model addresses parameter sharing, not pretrained scale or unseen-source inference. Gradient-norm control leaves direction, head adaptation and optimizer history coupled; the local model clarifies possible mechanisms without identifying a unique cause. A next test should use structured identities and previously unseen sources, followed by pretrained multi-token interfaces, with the same separation of identity access, gradient exposure and downstream compatibility.

AI USE STATEMENT

Large language models assisted with drafting and revising the manuscript, improving clarity and organization, and refining explanations and figure captions. Assistance also included literature retrieval, reference-code development and figure preparation. The authors retain responsibility for the final text, citations, technical claims and interpretation of results.

REPRODUCIBILITY STATEMENT

Appendix A specifies task generation, data splits, architectures, optimization, random seeds, gradient routes and readout fitting. Appendix B.1 defines paired comparisons and uncertainty estimation, and Appendix C provides the diagnostic derivations.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644v4. Version 4, revised 2018.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021. URL https://papers.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gKLAAfiytI.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1807.03819v3.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SwIp410B6aQ.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41340–41391, 2025. doi: 10.52202/085713-1380. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b01972cf31e6fa0fe29e4b8b5c2a0a1-Abstract-Conference.html.

Kaiyu He, Mian Zhang, Peilin Wu, Xinya Du, and Zhiyu Zoey Chen. Is grokking worthwhile? Functional analysis and transferability of generalization circuits in transformers. *arXiv preprint arXiv:2601.09049*, 2026. URL https://arxiv.org/abs/2601.09049v1. Version 1.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022. URL https://arxiv.org/abs/2202.08384v1. Version 1.

Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.

8845–8855, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Islam_A_Broad_Study_on_the_Transferability_of_Visual_Representations_With_ICCV_2021_paper.pdf.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3301–3310, 2019. URL https://proceedings.mlr.press/v97/kaya19a.html.

Maksim A. Kazanskii. How to tame grokking: Representation geometry as a control signal. *arXiv preprint arXiv:2607.11666*, 2026. URL https://arxiv.org/abs/2607.11666v1. Version 1.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2604.07822v2. Version 2.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28648–28662, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f0bf4a2da952528910047c31b6c2e951-Abstract.html.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 562–570, 2015. URL https://proceedings.mlr.press/v38/lee15a.html.

Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17710–17722, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/94130ea17023c4837f0dcdda95034b65-Abstract.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34651–34663,

2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfc310e81992d2e4cedc09ac47eff13e-Abstract-Conference.html.

Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Miłoś, and Tomasz Trzciński. The tunnel effect: Building data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76772–76805, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/f249db9ab5975586f36df46f8958c008-Abstract-Conference.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/pdf?id=9XFSbDPmdW.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177v1. Version 1.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2676109d49d1eb26d6bc584a8f556305-Abstract-Conference.html.

Till Speicher, Vedant Nanda, and Krishna P. Gummadi. Understanding the role of invariance in transfer learning. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=spJI4LSPIU.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95238–95265, 2024. doi: 10.52202/079017-3017. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad217e0c7fecc71bdf48660ad6714b07-Abstract-Conference.html.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27223–27250, 2023. doi: 10.52202/075280-1186. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56cbfbf49937a0873d451343ddc8c57d-Abstract-Conference.html.

A Experimental specification

A.1 Architecture and source populations

For sequence state H, the causal pre-RMSNorm block is

$$U = H + W_O \,\mathrm{MHA}(\mathrm{RMS}_1(H)), \tag{6}$$

$$F_\theta(H) = U + W_2 \,\mathrm{GELU}(W_1 \,\mathrm{RMS}_2(U)). \tag{7}$$

Attention has four heads of width $d/4$ and includes the current position. RMSNorm has learned gains, no mean subtraction and stabilizer ϵ_{norm}; GELU uses the exact erf definition. All native projections are bias-free. Token and position embeddings start independently standard-normal, gains start at one, and a projection with fan-in n starts uniformly on $[-1/\sqrt{n}, 1/\sqrt{n}]$. There are no loop-index embeddings, extra inter-loop norms, reinjection or caches across examples. Numerical settings are $\epsilon_{\mathrm{norm}} = 10^{-8}$, distance-hinge $\epsilon = 10^{-8}$, and unit-vector/scale threshold $\epsilon_{\mathrm{scale}} = 10^{-12}$.

Embed the two source tokens at positions zero and one, call the block twice and retain position one. Processing makes two one-position calls without a further position embedding. With one position, the attention softmax is identically one; this stage is a residual MLP with a normalized linear attention path. Append the reader at position two and make eight two-position consumer calls. Atomic queries omit the reader and keep the one-position eight-call tail. Final RMSNorm and the native head read the last position.

A block has $12d^2 + 2d$ parameters. With U_b distinct blocks, the model has $(V_{\mathrm{in}} + V_{\mathrm{out}} + 4)d + U_b(12d^2 + 2d)$ parameters, including three position embeddings and final RMS gain. Graphs use $(V_{\mathrm{in}}, V_{\mathrm{out}}, d) = (551, 500, 128)$; addition uses $(258, 257, 256)$. PAD is reserved. The shared models contain 331,904 and 919,808 parameters. Distinct-stage models use three blocks, allocated 2/2/8 calls, with 725,632 and 2,493,696 parameters. The twelve-block model instead assigns a distinct block to every call. It has 2,497,408 graph parameters or 9,576,192 addition parameters. For both unshared layouts, initialize all blocks by copying one sampled block, then give each copy separate gradient and optimizer storage. Initial functions and call counts therefore match the shared model; parameter counts and subsequent updates differ.

Task generation and legal pairs. Graph task seeds are 8000–8015. In head order, sample ten outgoing labels without replacement from 50 and targets independently with replacement from 500 entities. Select 250 designated facts by permutation. Enumerate two-hop paths in first-fact order and the bridge's outgoing order. Exclude designated-second paths; permute ordinary–ordinary paths and use the first $\lfloor 0.95N \rfloor$ for training. All 5,000 atomic facts receive supervision. The role-held-out population has 2,350–2,395 queries before the transfer eligibility restriction.

For each bridge, use NumPy seed $900000 + s$ to permute its ordinary outgoing facts, consuming the stream in bridge order, and assign the first half to the auxiliary reader set. An ordinary-source pair is eligible if its bridge identities agree and at least two assigned readers yield training compositions for both endpoints. Sample pairs uniformly with replacement and two compatible readers without replacement. Thus every forwarded auxiliary composition is a training row. The native CE stream still includes assigned and unassigned readers.

Addition task seeds are 8200–8215. For each residue m, sample eight distinct first operands a by permutation and set $c = (m - a) \bmod 257$. The first four decompositions are A, the remaining four B. Independently permute ordered (m, b) pairs and split at $\lfloor 257^2/2 \rfloor$ before source expansion. Commuted pairs may fall on different sides. Auxiliary batches draw uniformly from the $257\binom{4}{2} = 1{,}542$ A-source pairs, followed by two distinct training readers b. B sources and designated graph sources never enter an auxiliary pair; their original atomic labels remain trained. Original-task endpoints at 200k use graph role-held-out composition or addition queries that jointly use B sources and held-out canonical (m, b) pairs. The separate 5k in-role composition diagnostic uses a different split and is labeled accordingly.

A.2 Objectives, gradient routes and coverage

Shuffle the atomic-plus-composition training union once per epoch, with batches of 512 and a retained final short batch. Sum the two route-specific CE totals and divide by the actual batch size;

do not oversample atomics. FP32 AdamW uses learning rate 10^{-3} after a 50-update linear warmup, betas $(0.9, 0.98)$, epsilon 10^{-8} and global gradient clipping at one. Decay all parameters by 0.1 for graphs and 1 for addition. Checkpoints are fixed by update count. For task s and source initialization i, native and auxiliary streams use $100000 + 100s + i$ and $200000 + 100s + i$. Clone parameters, AdamW state and native-stream position at 5k; every branch replays the same native batches.

For fixed-coefficient supervision, reserve 256 native rows and 64 legal source pairs at the parent. With native and auxiliary source-parameter gradients g_N, g_A, set

$$\lambda = \gamma \frac{\|g_N\|_2}{\|g_A\|_2}. \tag{8}$$

Use $\gamma = 0.1$ unless a dose is stated. Absent gradients count as zero; zero or nonfinite norms fail calibration. Hold λ fixed over updates 5001–10000, remove the auxiliary loss afterward, and continue native CE to 200k. Native CE always differentiates all source parameters. A processing-only auxiliary pass detaches encoding and any downstream parameters while preserving derivatives with respect to their inputs. Calling the old consumer under `no_grad` would incorrectly remove this derivative. In the shared model, a P-route update can change future E and C forwards; the distinct-stage replication removes that parameter coupling.

Identity supervision. Atomic-CE averages the original atomic CE of all 128 endpoint occurrences, including repeats. The old eight-call consumer, final RMSNorm and identity head use detached parameters in the auxiliary pass. Direct-ID replaces that route by a new affine M-class head on raw w, adding $M(d+1)$ pulse-only parameters: 64,500 for graphs and 66,049 for addition. Its weights and biases use independent uniform initialization on $[-1/\sqrt{d}, 1/\sqrt{d}]$, with local seed $670000 + 100s + i$. Two equal-valued CE graphs separate the updates: detached head parameters give the calibrated P-gradient; detached states give the unscaled head gradient. The head's fresh AdamW uses learning rate 10^{-3}, betas $(0.9, 0.98)$, epsilon 10^{-8}, decay 0.1, a 50-update pulse-relative warmup and separate clipping at one. The source optimizer is inherited. Exclude head parameters from Equation 8 and discard the head before every transfer evaluation.

SupCon uses the same $N = 128$ endpoint occurrences and atomic identities. Define $v_i = w_i / \max(\|w_i\|_2, \epsilon_{\text{scale}})$ and $P(i) = \{p \neq i : m_p = m_i\}$. Each occurrence has a positive, including when a prefix is repeated. The objective is

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(v_i^\top v_p/\tau)}{\sum_{a \neq i} \exp(v_i^\top v_a/\tau)}. \tag{9}$$

All other occurrences, including positives, enter the denominator. The temperature is 0.1 and there is no projection head. Contraction uses the 64 sampled unordered pairs. SupCon instead includes every ordered pair of distinct endpoint occurrences with the same identity. If identity m occurs c_m times, it contributes $c_m(c_m-1)$ directed positive terms. With no repeated identity across the 64 sampled pairs there are 128 terms, or two directions per pair; repeats can raise this count. A repeated raw prefix still counts as an occurrence. Each anchor averages its positive terms, so counting terms does not by itself measure their gradient weight. In the first reserved auxiliary batch, averaged over sixteen tasks and four source initializations, SupCon has 169.75 directed positive terms on graphs (range 136–216) and 197.50 on addition (160–264). Across the 128 anchors, the denominators contain 16,256 directed non-self comparisons. These counts audit the reserved batch, not every intervention update. Initial norm calibration also does not match subsequent optimizer displacement.

Alternatives to semantic grouping. Consistency averages symmetric KL through the original consumer over both compatible readers, with encoding and consumer parameters detached on the auxiliary pass:

$$\mathcal{L}_F = \frac{1}{2|\mathcal{R}_{AB}|} \sum_{r \in \mathcal{R}_{AB}} [\mathrm{KL}(p_{A,r} \| p_{B,r}) + \mathrm{KL}(p_{B,r} \| p_{A,r})]. \tag{10}$$

Use temperature-one log-softmax with neither distribution detached. Isotropic substitutes $\mu \pm \delta_{AB} u/2$ into the Contraction hinge, with $\mu = (z_A + z_B)/2$ and uniform unit direction u from seed $250000 + 100s + i$. It penalizes random secants at the same input-distance scale. Shuffled uses different-identity endpoints matched by parent input distance and norm; the fixed matching procedure is given below.

Table 2: Evaluation coverage in each task family. A task is one independently generated graph or addition split. Source and consumer initializations are averaged within each task before uncertainty is computed. The nine-objective panel comprises Native, Atomic-CE, Direct-ID, Contraction, SupCon and the four alternative controls.

Comparison	Tasks and initializations	Readouts or measurements
Nine-objective panel	16 tasks; four source; three consumer	Fresh consumer, ridge and nearest view at 1/2/4 views; original task; geometry
Core five objectives	Same 16-task cohort	Retained computation at 0/2/4/8 calls; identity and factor readers at four views
Fixed and bounded budgets; depth, dose and timing	First eight tasks; four source; three consumer	Shared and twelve-block depth scans; regime comparisons; pulse diagnostics
Ridge regularization	All nine; full 16-task cohort	Five regularizers; 1/2/4 views; training and test accuracy
Identity and attribute probes	Core five; eight tasks; two source	Four-view source CV for both targets; one-view identity grid
Attribute-logistic sensitivity	Atomic-CE, Direct-ID and Contraction; same eight-task cohort	One-view regularizer grid
Geometry matching and reader exposure	Eight tasks; four source; three consumer where applicable	Matched identity/factor probes; four-view consumer at matched epoch exposure

GeomDR (Kazanskii, 2026) centers the 128 endpoint states after each processing call. With covariance $C_k = \widetilde{Z}_k^\top \widetilde{Z}_k/(N-1)$ and descending eigenvalues μ_j, its adapted spectral penalty is

$$\mathcal{L}_{\mathrm{GeomDR}} = \sum_{k\in\mathcal{P}} \sum_{j>r} \mu_j(C_k). \tag{11}$$

It uses centered SVDs, retains repeated endpoints and ignores their pair membership after pooling. The processing-call set is $\mathcal{P} = \{3, 4\}$, and rank $r = d/8$ is 16 or 32. This specifies an adaptation of the spectral objective, rather than a reproduction of its original full training system.

A.3 Supervision depth and gradient-budget controls

The supervision-depth experiment moves a fresh identity head along the original computation, while the retained-depth experiment below changes only the evaluation readout. At supervision depth $s \in \{0, 2, 4, 8\}$, apply the first s original atomic-consumer calls to w and attach the same newly initialized affine identity head, with no final RMSNorm. The encoding parameters and intervening consumer parameters are detached during the auxiliary pass, but the derivative through each consumer call reaches the processing stage. Native CE continues to differentiate the full model. Detaching parameters for this pass does not hold their forward values constant throughout training: native updates, and shared-weight processing updates, can change them.

All depths use the Direct-ID head initialization, optimizer and loss scaling. Depth zero is exactly Direct-ID. Depth eight has a fresh head, whereas Atomic-CE retains its original final normalization and trained head. The scan therefore holds head age and optimization fixed. Both the shared and twelve-block scans use the first eight tasks, four source initializations and three fresh-consumer initializations. The fixed-coefficient scan crosses supervision depth with $\gamma \in \{0.1, 0.5\}$; the bounded controller adds $\gamma = 0.1$ at every depth. In the twelve-block model, consumer and encoding parameters are distinct from the processing blocks, so processing updates cannot directly change those other parameter sets.

A bounded gradient budget. The common controller in Equation 4 applies to Atomic-CE, Direct-ID, Contraction and the supervision-depth heads. Initialize λ_0 with Equation 8 on the reserved parent batch. Each subsequent native/auxiliary minibatch supplies the full-source native gradient $g_{N,t}$ and processing-route auxiliary gradient $g_{A,t}$ before clipping. Use $\gamma = 0.1$ unless specified, $c_{\max} = 10$

and $\epsilon_g = 10^{-6}$, with detached norms and coefficient. Combine $g_{N,t} + \lambda_t g_{A,t}$, clip its norm at one, and apply the inherited source AdamW update. There is no running average or validation fitting. Initial norms must be positive and finite. A zero current auxiliary gradient remains zero; a zero native norm sets the coefficient to zero. For a positive native norm, the achieved pre-clipping auxiliary/full-native ratio is at most γ, and may be smaller when the cap binds.

Fresh heads retain the unscaled CE, detached-state route and separate AdamW settings specified in Appendix A.2. Auxiliary source losses detach head parameters; Atomic-CE and Contraction need no new head. The implementation checks optimizer ownership and the auxiliary parameter route before updating either optimizer.

Loss and gradient diagnostics. At each reported pulse checkpoint, retain auxiliary CE, raw auxiliary norm, full-source native norm, native norm restricted to processing parameters, coefficient, achieved ratio, cap status, native–auxiliary cosine and joint clipping factor. Norms and coefficients refer to the same update's batches. Zero-norm cosines are undefined. The processing-native norm is descriptive: Equation 4 uses the full native norm in every architecture. These diagnostics distinguish residual error from overall gradient magnitude; they do not identify which gradient directions benefit transfer. Separate backward evaluations also add computation.

Supervision exposure. A full 512-row native batch contains approximately 53.48 graph or 7.85 addition atomic occurrences in expectation; each auxiliary update adds 128 targeted endpoint occurrences. Short final batches slightly change total exposure. Occurrence counts do not equate gradient weights: norm calibration, processing-only routing and repeated endpoints all matter. Appendix B.3 reports the parent checkpoint's atomic and composition accuracies.

The fixed-coefficient loss-shape scans vary Direct-ID and Contraction over $\gamma \in \{0.05, 0.1, 0.2, 0.5\}$. Independently vary the Contraction hinge target over $\rho \in \{0.5, 0.7, 0.9\}$ at $\gamma = 0.1$, or the SupCon temperature over $\tau \in \{0.05, 0.1, 0.2\}$. Each setting retains the same source examples and random streams but recalibrates its own auxiliary coefficient at the parent. Report the full declared grid without selecting a setting on held-source accuracy. These scans compare initial gradient scale; they do not equate cumulative gradients or parameter displacement.

Intervention schedules and model depth. The schedule comparison uses auxiliary updates 5001–10000, 40001–45000, 160001–165000, or 5001–200000. The first three windows each last 5,000 updates; the sustained setting lasts 195,000. Each branch starts from its own native checkpoint immediately before the window, inherits the source optimizer state and calibrates its coefficient there. A fresh identity head, when required, restarts its 50-update warmup at the beginning of that window. All branches are evaluated at source update 200k. Thus later interventions also have less native training after the intervention, and sustained supervision has greater total exposure and cost. The schedule comparison does not isolate those factors.

The twelve-block comparison uses the same task distributions, width, E2/P2/C8 call allocation and optimizer as the shared source. Its processing-only auxiliary gradient reaches the two processing blocks, while native CE updates all twelve. The fixed-coefficient comparison includes Atomic-CE, Direct-ID, Contraction and SupCon; the bounded-budget comparison uses the first three objectives on the same eight-task cohort. The three-block and alternative-decay comparisons use their own native 5k parents, inherited optimizer states and norm calibrations, with the same cohort and downstream evaluation. Alternative decay is zero for graphs and 0.1 for addition. These decay values are reference settings, not demonstrated optima.

A.4 Frozen-source populations and readout fitting

Table 2 specifies the cohorts; source initialization indices start at zero. Four fresh balanced maps from all $M = 500$ entities or 257 residues to 16 classes use PyTorch seed $430000 + s$, in reader order. Assign labels $\lfloor 16m/M \rfloor$ after a random permutation of identities. Balance is over the full identity universe, not the graph subset selected below.

Select sources without model outputs. In graph identity order, use generator $420000 + s$ to permute the ordinary incoming sources of each identity having at least five ordinary and one designated source. Reserve the first ordinary source, use the next one/two/four as nested training views, and test

all designated sources. This retains 174–198 identities, 916–976 designated and 696–792 reserved-ordinary reader queries. For addition, permute the four A sources of each residue under the same seed rule, use the first one/two/four, and test all B sources (4,112 reader queries). Expand each selected prefix across four readers. Every tested identity–reader label appears in training at every budget, but no test prefix does.

At 200k cache w without gradients. For each task, source arm, initialization and budget, divide train and test states by the scalar training RMS $a = \sqrt{\mathrm{mean}_{x \in \mathrm{train},j}\, w_j(x)^2}$. Reject nonfinite a or $a \leq \epsilon_{\mathrm{scale}}$. Repeating each source across readers does not change a. No test statistic or per-coordinate scale is fitted.

Fresh consumer. Initialize four reader embeddings, one position vector, one causal block, final RMSNorm and a bias-free 16-way head, independently of the source model and identically across source arms at a given task/initialization/budget. Width and initialization distributions match the source model. Stack the normalized state and reader-plus-position embedding, make eight block calls, and read the final reader position. Three consumer initializations accompany each source initialization, with seed $650000 + 100s + 10i + j$, $j \in \{0, 1, 2\}$; minibatch shuffling uses $660000 + 100s + 10i + j$. Train for 2,000 CE updates with shuffled epochs, batches of 128 and retained short batches. Use AdamW learning rate $10^{-3} \min(u/50, 1)$, betas $(0.9, 0.98)$, epsilon 10^{-8}, decay 0.1 on all parameters and clipping at one. Record complete-population results every 100 updates including zero, and use the fixed 2,000-update endpoint for the primary comparison. The training population grows with the number of views, so this equal-update comparison does not give equal epoch exposure.

Attribute ridge probe. Fit four separate 16-output affine classifiers to the same normalized states and observed labels. For a reader with n examples, minimize

$$\frac{1}{n}\|XU + \mathbf{1}b^\top - Y\|_F^2 + \alpha\|U\|_F^2, \tag{12}$$

where Y is one-hot and the bias is unpenalized. Fix $\alpha = 0.01$ for all settings. Center X, Y using training means, solve $(X_c^\top X_c + n\alpha I)U = X_c^\top Y_c$ in double precision, and recover $b = \bar{Y} - U^\top \bar{X}$. Predict by argmax, breaking exact ties by class index. This deterministic fit has no consumer seed. Training and held-source accuracy are reported for every objective and view budget. With one view, the number of sources is close to the feature width, so a near-perfect training fit can coexist with unstable held-source predictions. Arbitrary attribute partitions also need not be linearly separable at that width. These are distinct limitations of the probe.

Probe robustness and model selection. Refit ridge probes at $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ for all nine objectives, all sixteen tasks, four source initializations and every view budget. The test labels do not select α; complete training and test arrays for this grid accompany the supplement.

The one-view logistic regularizer check uses the first eight tasks and first two source initializations for Atomic-CE, Direct-ID and Contraction. Four-view logistic CV also includes Native and SupCon on those same task and initialization pairs. For each of the four readers, minimize mean cross-entropy plus $\alpha\|U\|_F^2/2$, leaving the bias unpenalized. Use a training-only scalar RMS, zero parameter initialization and double-precision L-BFGS. The implementation checks the final gradient as well as the optimizer return status; its maximum permitted absolute gradient is 10^{-6}. Report one-view training and held-source accuracy across the same fixed regularizer grid.

At four views, select the logistic regularizer by cross-validation within the downstream training sources. Sort source IDs within each identity, hold out one source per identity in each of four folds, and move all four reader labels for a source together. Fit normalization and the head on the remaining three sources per identity. Select the regularizer with the smallest mean validation CE, choosing the larger value on numerical ties, then refit all four views. No test source enters normalization, fitting or selection. One-view training cannot support this procedure: withholding its only source would also remove that identity's complete attribute assignment. The one-view result therefore remains a regularizer sweep, not a cross-validated estimate.

Posthoc linear identity probes. Fit an affine ridge classifier to identity one-hot vectors at the 200k source checkpoint. The graph head has 500 outputs and the addition head 257. Graph fitting and

testing cover the same selected 174–198 identities; this does not test unobserved identity classes, and $1/500$ is not its chance reference. Use Equation 12, with the same training-only RMS and centering, unpenalized bias, double-precision solver and five-value regularizer grid. For unobserved graph output columns the centered target is zero, so the fitted column remains zero. Predictions use the full output vector.

Use the five-objective, eight-task, two-source cohort in Table 2 and report the complete one-view grid. At four views, form the source-held validation folds described above and refit normalization and centering within each fold. Select the penalty by mean *summed squared error* over the identity output vector, not CE on unconstrained ridge scores. Resolve validation-loss differences within 10^{-12} toward the larger penalty, then refit all four views. Every identity-probe cell selects $\alpha = 1$, the upper end of the declared grid; this is finite-grid selection, not evidence of a globally optimal penalty.

These heads are fitted after source training. Their accuracy measures endpoint linear identity readability; it does not show that the temporary head discarded at 10k would still decode the 200k state. Fresh-consumer endpoints average three reader initializations per source initialization; deterministic probes have no such replication.

Matching reader training exposure. Let M_{tr} be the number of training identities and let $B_v = \lceil 4vM_{\mathrm{tr}}/128 \rceil$ be the number of minibatches in an epoch with v views. The exposure comparison evaluates the four-view consumer at $U_4 = \mathrm{round}(2000B_4/B_1)$ updates, using the original shuffle and optimizer schedule. The one-view reference remains at 2,000 updates. This matches epoch exposure to within half a minibatch rather than matching update count. Every saved four-view evaluation up to 2,000 updates is retained. Longer curves show whether the original endpoint concealed subsequent learning or deterioration; they are not used to choose a favorable stopping time.

Nearest view and retained computation. Deduplicate the four reader rows by original source ID and sort training states by that ID. Find the single closest state by direct double-precision Euclidean distance and return its observed label for the requested reader. All four readers retrieve the same source; ties choose the lowest source ID. The prediction routine receives no test identity or target. A common positive RMS scale does not change the neighbor.

For retained depth $k \in \{0, 2, 4, 8\}$, apply the first k frozen original atomic-consumer calls to the *raw* state w. Do not append a reader, apply the native final RMSNorm or reuse the old identity head at an earlier exit. Fit a new training-only scalar RMS and the same affine-probe family independently at each depth. This measures accessibility along the old computation at selected exits; it does not estimate a minimum sufficient depth.

The complete retained decoder is a different control: apply all eight old calls, final RMSNorm and identity head, then look up the predicted identity's new label. Build that table only from downstream training prefixes, their original atomic-training identities and their observed new labels. An unknown identity abstains and scores as incorrect. The table is complete for all tested identities at one view, so its predictions do not vary with view budget. The old consumer uses the source model's frozen block in the shared layout; retaining it does not require a separately stored backbone.

A.5 Readers that conflict with source invariance

Assign four independent balanced 16-way maps to the graph's first relation $r_1 \in \{0, \ldots, 49\}$ or addition's first operand $a \in \mathbb{Z}_{257}$. The map generator uses seed $630000 + s$, in reader order. Replace the downstream targets by these factor labels, keeping the source state and reader index unchanged. Use the primary source split and the same affine fitting rule. The reported comparison uses four views, with the identity reader also rescored on the identical filtered test set. Thus an objective-by-reader contrast changes the requested labels, not the test sources or fitting budget. The labels can now differ between sources of the same intermediate identity.

Some factor categories may be absent from one-view fitting. Fix test eligibility using the categories observed in one-view training, then retain exactly that test set for every objective and all three budgets. Across tasks, this retains 880–968 graph reader queries (90.91–100% of the primary held-source queries) and 2,444–2,788 addition queries (59.44–67.80%). This conditions the test on known factor-label assignments, rather than rewarding chance guesses for unseen inputs to a ran-

dom map. It differs from filtering only on output-class coverage. The intermediate-identity lookup cannot generally answer these readers, since an identity can carry multiple factor labels.

A.6 Checkpoint geometry, matching and replications

At 200k take the union of the four selected training sources and all held-source test prefixes, once per original source ID. Let μ_m be each identity's state centroid and $\bar{\mu}$ the equally weighted mean of those centroids. Compute

$$W = \frac{1}{M}\sum_m \frac{1}{n_m}\sum_{s:m_s=m} \|w_s - \mu_m\|_2^2, \qquad B = \frac{1}{M}\sum_m \|\mu_m - \bar{\mu}\|_2^2. \tag{13}$$

Here M is the number of retained identities. Report W and B divided by the squared one-view training RMS, alongside W/B: uniform state scaling leaves all three reported quantities unchanged. Test identities are used only for this diagnostic, never to fit a reader or choose a neighbor. Identity-neighbor retrieval scores whether the nearest training source has the query's intermediate identity. Appendix C relates this diagnostic to random-attribute nearest-view accuracy.

Geometry association and matching. To assess whether separation predicts transfer, first average source and consumer initializations within each task. For each task separately, compute Pearson and Spearman correlations across the nine fixed objectives between fresh-consumer accuracy and either $\log(W/B)$ or identity-neighbor accuracy. Average the task correlations on the Fisher-z scale and form the interval across the sixteen tasks before transforming back. The nine objectives are fixed comparisons, not nine independent experiments. A descriptive straight-line fit of accuracy on $\log(W/B)$ omits SupCon; its prediction error on SupCon exposes where the aggregate association is insufficient. It is not a validated predictor for a new objective.

The compatibility sensitivity compares Direct-ID at its default setting with the Contraction grid over gradient fraction and hinge target. For each task and source initialization separately, choose the Contraction setting whose $\log(W/B)$ on the four training views is closest to Direct-ID's. Break exact ties by lexicographic setting order, without consulting held-source labels. Report the remaining geometry mismatch and evaluate identity and factor readers on the same factor-covered test sources. This is a comparison among discrete settings; equal scalar scatter ratios do not equate covariance spectra, directions of variation or all aspects of a representation.

For Shuffled, draw 4,096 legal training pairs with seed $230000 + 100s + i$. Bin their parent input distances and mean endpoint norms using linear-interpolated quantiles at $0, 0.2, 0.4, 0.6, 0.8, 1$, including underflow and overflow bins. Match each first endpoint to a second endpoint of a different identity whose original pair shares its bin. Use maximum-cardinality one-to-one matching with lexicographic priorities and breadth-first augmenting paths. Freeze this pool and sample 64 pairs with replacement per pulse update; do not re-bin the crossed pairs.

Computation and reference implementation. A block on L tokens uses $12Ld^2 + 2L^2d$ dense multiply–accumulates (MACs), including the one-token query and key projections. A source composition query therefore uses 4,400,128 MACs for graphs and 17,388,800 for addition. Parameter sharing changes storage, not these call counts. The fresh identity head adds 64,500 or 66,049 parameters only while auxiliary supervision is active. For supervision depths zero and eight, the auxiliary forward cost per endpoint is respectively 1,246,208 and 2,821,120 MACs on graphs, or 4,789,504 and 11,085,056 on addition. These analytic counts exclude nonlinearities, backward passes and optimizer operations; they are not hardware runtime measurements. The supplement includes the source architecture, task generation, gradient routes, probe solvers and deterministic figure builders.

B Supplementary results and inference

B.1 Pairing and uncertainty

Pair source objectives within task and source initialization, and within consumer initialization where applicable. For four-source-initialization comparisons, average all twelve fresh-consumer differences (four source times three consumer initializations) within each task. The common-cohort main

table instead averages six (two source times three consumer initializations). Deterministic readouts average four or two differences according to the stated source cohort. With task means $\bar{\Delta}_j$, report

$$\bar{\Delta} \pm t_{J-1,0.975} \frac{s_{\text{task}}}{\sqrt{J}}. \tag{14}$$

The complete nine-objective panel has $J = 16$ tasks. The common-cohort main table and intervention comparisons have eight. These intervals describe task variation conditional on the selected source initializations. Queries, checkpoints and repeated views are not independent samples. Across the four source initializations, the one-view Direct-ID minus Atomic-CE task means range from 4.14 to 5.87 points on graphs and 6.99 to 8.02 on addition. The corresponding Contraction ranges are 7.76 to 8.44 and 8.75 to 9.01. These means describe the selected initializations; they do not estimate variation over all possible initializations.

The four fixed-coefficient baseline hypotheses at one view are Direct-ID minus Atomic-CE and Contraction minus Atomic-CE in each task family. Use two-sided paired t tests with Holm correction at familywise level 0.05; displayed intervals remain pointwise. All other contrasts, including new controller and depth comparisons, SupCon, geometry and conflicting readers, are descriptive. For an objective-by-controller interaction, take the objective-minus-Atomic contrast under the bounded rule and subtract the corresponding fixed-coefficient contrast within each task. For a regime interaction, subtract the shared-setting objective gap from the alternative-setting gap within each task before forming the interval. Whether two separate intervals cross zero does not test that interaction. A confidence interval crossing zero is not an equivalence result. For budget interactions, compute each task's change in an objective gap from one to four views under each readout, then subtract these changes before forming the interval.

Why the factor-reader intervals are narrow. For paired task scores A and B, the difference variance is

$$\mathrm{Var}(A - B) = \mathrm{Var}(A) + \mathrm{Var}(B) - 2\,\mathrm{Cov}(A, B). \tag{15}$$

On graphs, the task standard deviations of Contraction and Atomic-CE factor accuracy are 1.62 and 1.45 points, with correlation 0.82. Their paired difference therefore has standard deviation 0.94 points, giving a 95% interval half-width of 0.50 over sixteen tasks. On addition, the corresponding standard deviations are 0.53 and 0.70, the correlation is 0.61, and the difference standard deviation is 0.57, giving a half-width of 0.30. These are task-level calculations, not binomial intervals over thousands of reader queries. Pairing reduces variation, but the narrow intervals also reflect the limited factor variation in these task populations. The intervals should therefore be read as conditional on this task family, not as estimates of the cost for arbitrary source factors.

B.2 Readout endpoints and the transfer checkpoint

Table 4 reports all nine objectives with affine training fit at every view budget. Table 5 aligns original-task performance and geometry to the 200k transfer checkpoint; Table 6 compares retained readout depths. Table 7 gives the full linear identity-probe grid for the five fixed-coefficient objectives in the common-cohort main comparison.

B.3 Supervision controls and training diagnostics

Table 8 reports Contraction under both coefficient rules, complementing the Direct-ID comparison in the main text. Table 9 crosses supervision depth with dose and parameter sharing. Depth zero reuses Direct-ID; the depth-eight fresh-head comparison remains distinct from Atomic-CE. These comparisons retain the same native update count and source examples; later windows and continuous supervision have different cumulative auxiliary exposure.

Loss, gradient norm and coefficient clipping. Figure 4 shows the auxiliary CE and raw-gradient profiles. Table 10 integrates the achieved auxiliary/full-native norm ratio over all 5,000 auxiliary updates, excluding the initial calibration point. In the shared graph model, the bounded rule attains mean ratios 0.088 at depth zero and 0.059 at depth eight. Their coefficient caps bind on 32.7% and 70.3% of updates, respectively. Thus a common upper budget does not give equal realized pressure. Any remaining depth difference can still reflect unequal magnitude, gradient direction and optimizer history.

Table 3: A common-cohort readout comparison of the fixed-coefficient source objectives at 200k. Accuracy (%) over the same eight tasks and two source initializations; fresh consumers average three additional reader initializations. CV uses only the four selected training sources per identity. Original-task accuracy uses role-held-out composition for graphs and the joint B-source / held-canonical composition test for addition.

Task	Objective	Fresh: 1	Fresh: 4	Logistic CV: 4	Identity CV: 4	Original
Graph	Native	45.12	77.73	26.48	73.64	59.14
	Atomic-CE	63.77	86.18	32.42	89.68	65.47
	Direct-ID	67.67	88.07	44.08	99.35	67.09
	Contraction	72.27	91.06	45.16	99.35	66.72
	SupCon	69.99	88.66	46.51	99.32	67.33
Addition	Native	50.06	82.01	28.92	90.03	78.31
	Atomic-CE	66.44	90.88	36.99	97.90	89.05
	Direct-ID	75.11	93.07	52.73	99.99	88.94
	Contraction	78.19	95.42	54.47	100.00	89.19
	SupCon	73.26	92.80	53.64	99.96	89.43

Table 4: Complete source-view results at 200k. Entries are percentages at one / two / four views. Each task averages four source initializations; fresh consumers additionally average three reader initializations. Ridge uses the fixed baseline penalty $\alpha = 0.01$. All nine objectives and all fitting budgets are shown.

Task	Objective	Fresh consumer	Ridge test	Ridge fit
Graph	Native	44.4 / 61.5 / 77.5	12.4 / 18.5 / 24.4	99.9 / 88.2 / 67.7
	Atomic-CE	63.3 / 74.3 / 84.5	14.4 / 23.0 / 31.2	99.9 / 89.6 / 71.8
	Direct-ID	68.7 / 81.1 / 89.3	18.8 / 31.2 / 41.5	99.9 / 91.6 / 77.3
	Contraction	71.4 / 82.1 / 89.8	18.6 / 32.4 / 42.8	99.9 / 91.7 / 78.0
	SupCon	70.3 / 81.1 / 88.8	19.3 / 32.7 / 43.5	99.9 / 91.7 / 78.3
	Consistency	64.2 / 75.3 / 84.9	14.5 / 23.2 / 31.3	99.9 / 89.5 / 71.9
	Isotropic	44.7 / 61.2 / 77.7	12.1 / 17.8 / 23.5	99.9 / 88.0 / 67.0
	Shuffled	40.8 / 58.2 / 74.2	10.9 / 15.6 / 20.4	99.9 / 87.2 / 64.8
	GeomDR	65.0 / 76.7 / 86.8	17.4 / 28.1 / 37.8	99.9 / 90.9 / 75.8
Addition	Native	47.7 / 65.9 / 81.4	10.2 / 16.8 / 24.9	100.0 / 98.6 / 83.4
	Atomic-CE	68.2 / 80.1 / 91.6	11.9 / 22.2 / 33.5	100.0 / 99.0 / 87.1
	Direct-ID	75.8 / 86.2 / 93.4	15.7 / 33.9 / 50.6	100.0 / 99.4 / 92.5
	Contraction	77.0 / 87.6 / 94.3	15.9 / 33.7 / 50.3	100.0 / 99.4 / 92.4
	SupCon	74.4 / 86.2 / 93.4	15.8 / 33.6 / 50.1	100.0 / 99.4 / 92.3
	Consistency	69.1 / 80.9 / 90.8	12.3 / 23.3 / 35.3	100.0 / 99.1 / 87.8
	Isotropic	48.6 / 65.9 / 82.6	10.4 / 17.6 / 26.4	100.0 / 98.7 / 84.1
	Shuffled	44.7 / 63.5 / 80.8	9.8 / 16.0 / 23.5	100.0 / 98.6 / 82.8
	GeomDR	70.4 / 82.1 / 94.1	13.8 / 27.5 / 41.9	100.0 / 99.2 / 89.9

Reader exposure. Table 11 compares the fixed-update and approximately matched-epoch reader endpoints. This check addresses reader fitting time, whereas the supervision schedules above change source training. The two forms of exposure should not be conflated.

Intervention-start accuracies. At 5k, full-atomic accuracy is $70.05\,[68.39, 71.70]\%$ on graphs and $75.77\,[74.23, 77.31]\%$ on addition (sixteen tasks, four source initializations averaged within task). These are accuracies of the original decoder on its trained atomic population, not new heads or held-source probes. Composition diagnostics at the same update use sixteen tasks and two source initializations: graph composition-training accuracy is 36.81% with interval $[35.28, 38.35]$, and addition training accuracy is 36.86% $[35.16, 38.56]$. The corresponding in-role held-out accuracies are 4.98% $[4.67, 5.28]$ for ordinary-first/ordinary-second graph paths and 5.27% $[5.01, 5.54]$ for A-source addition with held-out canonical pairs. These differ in population and initialization cohort from the full-atomic check and the 200k source-task endpoints; they do not identify a particular grokking transition.

Table 5: Original-task accuracy and geometry at the same 200k checkpoint. Original-task accuracy uses role-held-out graph composition and, for addition, B sources with held-out canonical operand pairs; it is not the in-role test. Scatter components W, B use the one-view training RMS; their ratio and one-view identity-neighbor accuracy describe the frozen source bank. Sixteen tasks and four source initializations per family.

Task	Objective	Original (%)	W	B	W/B	NN identity (%)
Graph	Native	60.32	83.95	43.21	1.95	13.08
	Atomic-CE	67.74	78.52	48.52	1.63	25.94
	Direct-ID	68.63	70.21	56.99	1.24	58.53
	Contraction	67.98	67.29	59.83	1.13	75.77
	SupCon	68.29	66.18	61.29	1.09	79.60
	Consistency	67.83	78.59	48.66	1.63	27.85
	Isotropic	61.26	84.46	42.72	1.99	12.29
	Shuffled	59.16	86.87	40.38	2.16	9.92
	GeomDR	67.43	73.30	53.95	1.37	45.44
Addition	Native	78.47	186.98	68.61	2.74	24.66
	Atomic-CE	86.15	175.20	80.38	2.20	48.04
	Direct-ID	86.16	153.40	102.08	1.51	89.43
	Contraction	86.77	148.82	106.66	1.40	96.46
	SupCon	86.56	147.41	108.23	1.37	96.60
	Consistency	85.71	173.01	82.39	2.11	53.31
	Isotropic	78.99	185.45	70.37	2.65	26.56
	Shuffled	77.86	188.86	66.78	2.84	21.68
	GeomDR	84.88	164.22	91.44	1.80	73.69

Table 6: Readout after retained original computation. One-view affine attribute accuracy (%) after k frozen atomic-consumer calls; every exit refits the same ridge family. The original final norm and identity head are omitted. Sixteen tasks and four source initializations. This varies readout depth, separately from the supervision-depth experiment.

Task	Objective	$k=0$	$k=2$	$k=4$	$k=8$
Graph	Atomic-CE	14.43	32.20	56.69	88.62
	Direct-ID	18.76	41.42	63.73	90.13
	Contraction	18.55	41.49	63.05	89.23
	SupCon	19.32	42.23	63.46	89.70
Addition	Atomic-CE	11.91	27.02	59.96	99.24
	Direct-ID	15.73	38.35	72.00	99.58
	Contraction	15.93	38.57	71.00	99.41
	SupCon	15.80	38.71	72.14	99.52

Geometry between source checkpoints. Figure 5 follows the five fixed-coefficient objectives at 5k, 10k, 40k, 80k, 120k and 200k on unchanged identity and source populations. Contraction and SupCon can lose some early separation before recovering it later. These diagnostics concern the fixed-coefficient banks, not the adaptive-controller endpoints.

B.4 Geometry association and compatibility sensitivity

Figure 6 places all nine objectives on the separation–transfer axes. The task-averaged Pearson correlation of $\log(W/B)$ with fresh-consumer accuracy is -0.91 on graphs and -0.89 on addition; identity-neighbor accuracy gives 0.84 and 0.91. SupCon lies 6.02 and 7.48 points below the straight-line prediction fitted to the other eight objectives. This descriptive residual shows where the scalar predictor is insufficient. Neither its geometry difference nor its fresh-consumer difference from Contraction is resolved by the corresponding paired interval, so the result does not identify a SupCon-specific mechanism.

Table 7: Linear identity-probe sensitivity for fixed-coefficient source objectives on the same eight tasks and two source initializations. Entries give held-source / training accuracy (%). Graph heads have 500 outputs but only 174–198 observed training identities; addition uses all 257 identities. Four-view CV selects the squared-error penalty from training sources only. One-view columns show the complete fixed grid.

		One view; fixed ridge α					Four views
Task	Objective	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	Source CV
Graph	Native	9.6/100.0	9.8/100.0	11.1/100.0	16.6/100.0	22.4/100.0	73.6/100.0
	Atomic-CE	16.2/100.0	16.4/100.0	18.2/100.0	28.2/100.0	38.2/100.0	89.7/100.0
	Direct-ID	35.5/100.0	36.1/100.0	40.1/100.0	58.8/100.0	73.7/100.0	99.4/100.0
	Contraction	37.2/100.0	37.7/100.0	41.8/100.0	61.1/100.0	81.0/100.0	99.3/100.0
	SupCon	38.3/100.0	38.9/100.0	43.5/100.0	65.5/100.0	85.3/100.0	99.3/100.0
Addition	Native	1.3/100.0	2.4/100.0	6.6/100.0	18.6/100.0	32.1/100.0	90.0/100.0
	Atomic-CE	1.6/100.0	4.1/100.0	13.5/100.0	37.3/100.0	57.0/100.0	97.9/100.0
	Direct-ID	3.3/100.0	9.8/100.0	35.1/100.0	77.6/100.0	93.0/100.0	100.0/100.0
	Contraction	3.3/100.0	10.0/100.0	36.3/100.0	81.9/100.0	97.3/100.0	100.0/100.0
	SupCon	3.2/100.0	10.1/100.0	34.8/100.0	79.2/100.0	96.1/100.0	100.0/100.0

Table 8: Contraction minus Atomic-CE under each coefficient rule, using the same eight tasks, four source initializations and three reader initializations as the Direct-ID comparison. Entries are one-view fresh-consumer differences in percentage points with paired 95% task intervals.

Setting	Graph		Addition	
	Fixed	Bounded	Fixed	Bounded
Shared, 5k–10k	8.4 [5.8, 11.0]	8.0 [6.3, 9.7]	11.7 [5.8, 17.6]	7.8 [5.8, 9.8]
40k–45k	6.4 [2.9, 9.9]	7.0 [4.4, 9.6]	11.3 [4.8, 17.9]	10.1 [8.7, 11.5]
160k–165k	8.1 [4.6, 11.6]	5.9 [3.0, 8.8]	10.5 [3.5, 17.4]	9.6 [7.3, 11.8]
5k–200k	9.2 [4.2, 14.2]	10.8 [8.8, 12.7]	12.7 [6.3, 19.1]	9.2 [5.7, 12.8]
Three stages	6.7 [1.0, 12.4]	7.7 [5.4, 10.1]	7.9 [3.2, 12.7]	10.3 [7.4, 13.2]
Other decay	6.0 [1.9, 10.0]	5.4 [3.7, 7.2]	7.5 [1.5, 13.4]	9.1 [7.9, 10.3]
Twelve blocks	3.8 [-2.0, 9.7]	5.9 [3.5, 8.3]	7.3 [1.6, 12.9]	8.2 [6.6, 9.8]

Table 9: Supervision depth crossed with strength and weight sharing. Entries give one-view fresh-consumer accuracy (%); the final column gives the paired depth-zero minus depth-eight difference and its 95% interval. Every depth uses the same new-head family and evaluates the exposed state. Eight tasks, four source and three reader initializations.

Task/layout	Rule	$s = 0$	$s = 2$	$s = 4$	$s = 8$	$s = 0$ minus 8
Graph/shared	Fixed $\gamma = 0.1$	68.12	66.34	64.79	62.77	5.3 [3.9, 6.8]
Graph/shared	Fixed $\gamma = 0.5$	69.16	67.42	66.39	66.55	2.6 [0.8, 4.4]
Graph/shared	Adaptive $\gamma = 0.1$	70.41	70.17	69.46	68.50	1.9 [0.6, 3.2]
Graph/untied	Fixed $\gamma = 0.1$	68.59	67.54	66.66	65.60	3.0 [1.4, 4.5]
Graph/untied	Fixed $\gamma = 0.5$	69.18	69.04	68.39	67.89	1.3 [-1.2, 3.8]
Graph/untied	Adaptive $\gamma = 0.1$	73.09	72.20	71.37	72.03	1.1 [0.1, 2.0]
Addition/shared	Fixed $\gamma = 0.1$	75.18	73.14	69.83	69.77	5.4 [4.6, 6.2]
Addition/shared	Fixed $\gamma = 0.5$	76.40	74.21	72.51	73.90	2.5 [-0.3, 5.3]
Addition/shared	Adaptive $\gamma = 0.1$	73.84	73.93	72.29	71.83	2.0 [0.9, 3.1]
Addition/untied	Fixed $\gamma = 0.1$	78.62	77.67	75.51	74.90	3.7 [2.4, 5.1]
Addition/untied	Fixed $\gamma = 0.5$	79.34	79.35	77.06	77.72	1.6 [0.1, 3.1]
Addition/untied	Adaptive $\gamma = 0.1$	79.55	79.32	78.33	76.39	3.2 [1.8, 4.5]

Table 12 reports the training-geometry match and its held-source accuracy differences. The remaining scatter-ratio mismatch is shown explicitly; even an exact scalar match would not equate covariance orientation or source-factor content.

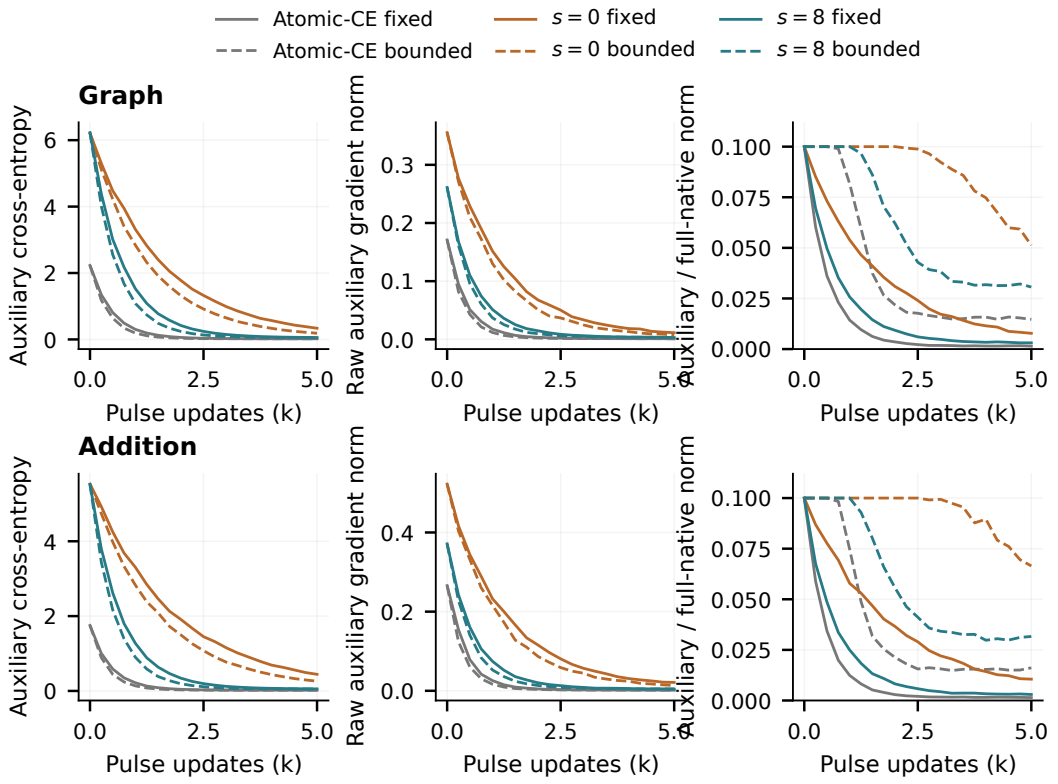

Figure 4: **Identity losses and gradient profiles during auxiliary supervision.** Rows show the shared graph and addition models. Columns show auxiliary CE, raw processing-gradient norm and attained auxiliary/full-native norm ratio. Colors distinguish original Atomic-CE from fresh heads at depths zero and eight; solid and dashed lines show fixed calibration and bounded control. Each task averages four source initializations before averaging the eight tasks. Other depths and untied models enter Table 10; its cap frequencies use all auxiliary updates.

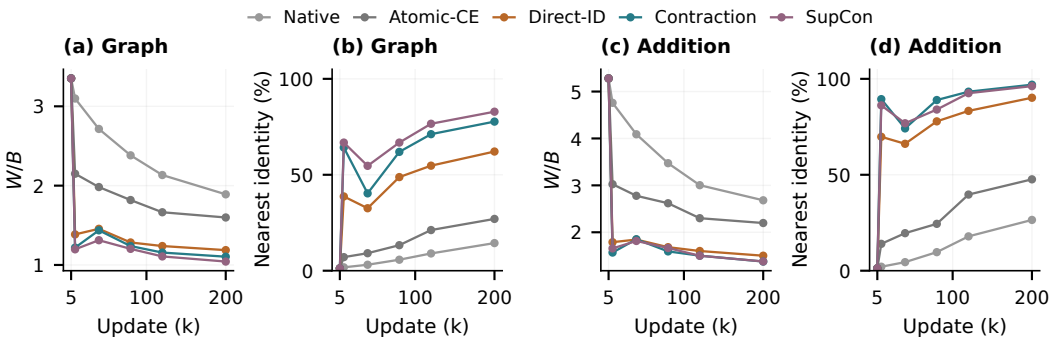

Figure 5: **Fixed-coefficient geometry across checkpoints.** Panels (a,b) show graphs and (c,d) addition, alternating W/B and one-view nearest-identity accuracy. Means use eight tasks and two source initializations. All objectives share the 5k parent and use the original 200k evaluation banks.

Table 10: Source-gradient budgeting over all pulse updates 1–5,000, excluding the calibration point. Entries average the attained pre-clipping auxiliary/full-native norm ratio; the final column is the adaptive coefficient cap frequency (%). These quantities are calculated per update before averaging source initializations and tasks.

Task/layout	Identity objective	Fixed 0.1	Fixed 0.5	Adaptive 0.1	Cap (%)
Graph/shared	Original Atomic-CE	0.011	–	0.038	80.9
Graph/shared	$s = 0$	0.033	0.123	0.088	32.7
Graph/shared	$s = 2$	0.028	0.103	0.081	45.9
Graph/shared	$s = 4$	0.023	0.086	0.071	56.2
Graph/shared	$s = 8$	0.017	0.063	0.059	70.3
Graph/untied	Original Atomic-CE	0.011	–	0.037	81.6
Graph/untied	$s = 0$	0.032	0.121	0.087	33.6
Graph/untied	$s = 2$	0.029	0.107	0.082	43.1
Graph/untied	$s = 4$	0.024	0.090	0.074	53.6
Graph/untied	$s = 8$	0.015	0.057	0.055	73.9
Addition/shared	Original Atomic-CE	0.010	–	0.036	82.2
Addition/shared	$s = 0$	0.037	0.140	0.094	21.7
Addition/shared	$s = 2$	0.028	0.105	0.081	44.0
Addition/shared	$s = 4$	0.022	0.082	0.070	58.4
Addition/shared	$s = 8$	0.016	0.061	0.057	71.5
Addition/untied	Original Atomic-CE	0.011	–	0.038	80.7
Addition/untied	$s = 0$	0.034	0.129	0.091	28.7
Addition/untied	$s = 2$	0.026	0.098	0.078	49.1
Addition/untied	$s = 4$	0.021	0.078	0.067	61.0
Addition/untied	$s = 8$	0.017	0.064	0.059	69.8

Table 11: Four-view reader exposure check on eight tasks. Match the number of shuffled epochs to the one-view 2,000-update budget, using $U_4 = \mathrm{round}(2000 B_4/B_1)$ with B_v batches per epoch. The rounding discrepancy is at most half a minibatch. All original curve points through 2,000 updates are unchanged.

Objective	Four-view endpoint	Graph	Addition
Atomic-CE	2,000 updates	85.70 [81.12, 90.27]	90.65 [87.57, 93.73]
Atomic-CE	Matched exposure	85.86 [80.85, 90.88]	91.72 [87.98, 95.47]
Atomic-CE	Additional-training change	0.17 [-0.74, 1.08]	1.08 [-0.18, 2.34]
Direct-ID	2,000 updates	88.62 [84.19, 93.05]	93.24 [89.84, 96.65]
Direct-ID	Matched exposure	89.07 [84.64, 93.50]	93.28 [89.60, 96.97]
Direct-ID	Additional-training change	0.45 [-0.68, 1.58]	0.04 [-0.79, 0.87]
Contraction	2,000 updates	90.96 [88.94, 92.99]	95.28 [92.76, 97.80]
Contraction	Matched exposure	90.64 [87.06, 94.22]	94.37 [91.79, 96.95]
Contraction	Additional-training change	-0.32 [-2.04, 1.39]	-0.91 [-1.66, -0.16]

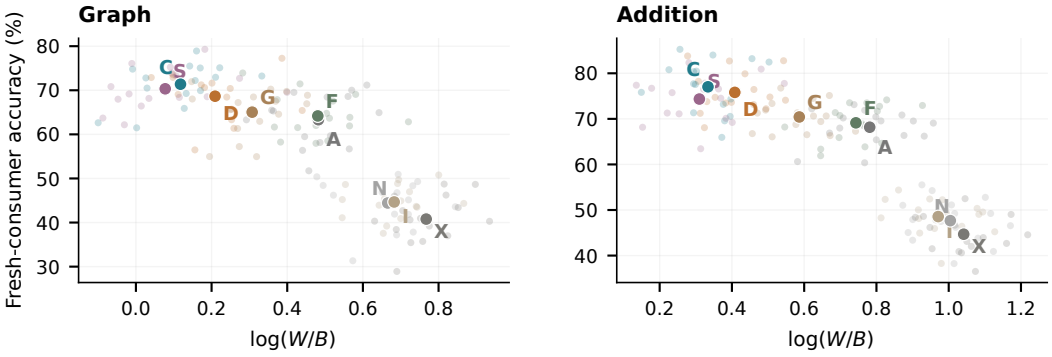

Figure 6: **Separation and fresh-consumer performance.** Faint points are task means; large points are objective means at 200k. N: Native; A: Atomic-CE; D: Direct-ID; C: Contraction; S: SupCon; F: Consistency; I: Isotropic; X: Shuffled; G: GeomDR. Correlations are within-task; the SupCon calibration omits that objective.

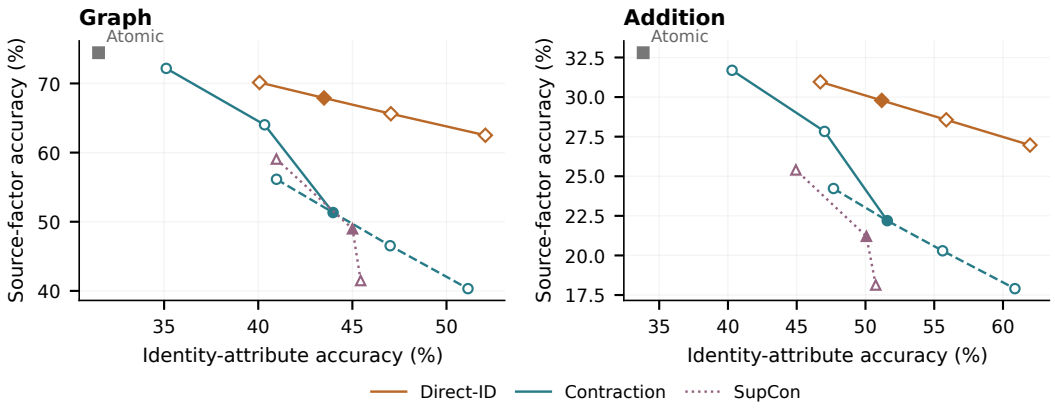

Figure 7: **Identity and source-factor reading.** All fixed-coefficient settings are shown: Direct-ID varies γ; Contraction varies ρ (solid) or γ (dashed); SupCon varies τ (dotted). Filled markers denote defaults. Both readers use four views on identical factor-covered sources; means use eight tasks and four source initializations. Matching intervals appear in Table 12.

Table 12: Training-geometry matching on eight tasks and four source initializations. Select the nearest Contraction grid point to default Direct-ID using only four-view training W/B. Identity-attribute/source-factor differences are selected Contraction minus default Direct-ID; log mismatch measures imperfect matching. Accuracy differences are percentage points. The finite grid does not make W/B identical and does not isolate a causal effect of geometry.

Check	Quantity	Graph	Addition
Geometry match	Identity attr. accuracy	-2.15 [-3.38, -0.92]	-1.50 [-5.19, 2.20]
Geometry match	Source-factor accuracy	-10.91 [-17.05, -4.76]	-5.36 [-8.21, -2.52]
Geometry match	Absolute log mismatch	0.03 [0.01, 0.04]	0.03 [0.01, 0.05]

C READING THE DIAGNOSTICS

These elementary calculations clarify the diagnostics; they are not a theory of the training outcomes.

Residuals, Jacobians and gradient direction. Equation 5 follows by differentiating cross-entropy through the affine head and consumer, then the processing call path. In a shared model this excludes explicit auxiliary derivatives through the consumer's shared parameters. It is not the full derivative of the ordinary source forward pass.

Consider binary cross-entropy with target one, $H_s(w) = \mathrm{diag}(q^s, 1)w$, $0 < q < 1$, and logit $(1,1)H_s(w)$. At $w = 0$, every depth has loss $\log 2$, residual $-1/2$, and gradient $g_s = -\frac{1}{2}(q^s, 1)$. Normalizing this gradient to unit norm leaves first-coordinate descent pressure

$$\frac{q^s}{\sqrt{1+q^{2s}}}. \tag{16}$$

For $q = 0.7$ and $s = 0, 2, 4, 8$, this is approximately $0.707, 0.440, 0.233, 0.058$. Equal loss and equal norm thus need not yield equal directional pressure. Conversely, the head $(q^{-s}, 1)$ gives the same gradient $-\frac{1}{2}(1,1)$ at every depth. Neither depth attenuation nor a transfer advantage follows without assumptions about the head and the directions useful to the later task. The supplement checks both calculations by automatic differentiation.

Why nearest-view attributes measure identity retrieval. Fix states and neighbor assignments before drawing a balanced attribute map. If p_{ID} is correct identity retrieval and its class sizes are $n_1, \ldots, n_{16}$ over M identities, then

$$q_M = \frac{\sum_c n_c(n_c-1)}{M(M-1)}, \qquad \mathbb{E}_f[\mathrm{Acc}_{\mathrm{nearest}}] = p_{\mathrm{ID}} + (1-p_{\mathrm{ID}})q_M. \tag{17}$$

Correct identity retrieval guarantees the attribute; otherwise q_M is the label-collision probability. Balancing precedes source eligibility, so $M = 500$ or 257 and $q_M \approx 0.06063$ or 0.05885. Four realized maps need not equal this expectation. It does not apply to a metric fitted to attribute labels.

Identity discrimination does not imply arbitrary linear attributes. The four states $(\pm 1, \pm 1)$ are linearly identifiable by using them as head weights: correct logits are two, other logits at most zero. Labeling opposite corners alike instead gives XOR, which no affine binary classifier separates. An identity decoder followed by a label lookup still solves this task.

Equal scalar separation need not retain the same factors. For each identity m, take four equally weighted sources $(f, n) \in \{-1, +1\}^2$ and centered prototypes μ_m. The banks

$$w^{(A)}_{m,f,n} = (\mu_m, f), \qquad w^{(B)}_{m,f,n} = (\mu_m, n)$$

have identical $W = 1$ and $B = \mathbb{E}_m\|\mu_m\|^2$. Factor f is linearly readable in A but balanced conditional on every B state, giving at most 50% factor accuracy there. Equal W/B therefore leaves factor retention unspecified; it does not imply that the compared experimental banks have otherwise equal geometry.