WHEN REPRESENTATION EFFECTS DEPEND ON THE MEASUREMENT CONDITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation analysis frequently treats an observed quantity as an intrinsic property of the representation being analyzed, despite the fact that it is generated by a particular measurement procedure. We make this dependency explicit by writing $R = \mathcal{M}(F; C)$, where F denotes the representation being measured and C denotes the measurement condition. This leads to a concrete invariance question: with F held constant, does the measured size of a representational contrast stay the same as C varies? To examine this empirically, we design a controlled protocol that keeps the representation fed into the measurement instrument bitwise identical while changing only the measurement condition. For vision representations, we find that two trained conditions yield very similar pairwise structure, yet they can assign markedly different magnitudes to identical representational contrasts. Therefore, matching relational structure does not guarantee that magnitudes transfer across conditions. Operating-point normalization accounts for a substantial portion of the observed scale gap, but does not eliminate it, and simpler probe families show the same qualitative behavior. Relative ordering and spatial concordance are far more stable than absolute magnitude, although they are not strictly invariant and may become effectively unmeasurable under a degenerate measurement condition. These findings support a measurement-conditioned view of representation analysis: a measured representational quantity should be attributed to F alone only when invariance across the relevant measurement conditions is demonstrated; otherwise, the condition should remain part of the claim's scope, $X(F; C)$.

1 INTRODUCTION

1.1 THE IMPLICIT MEASUREMENT CONDITION

Representation analysis often makes statements of the form "the representation F has property X." Such a property is rarely observed directly; instead, we infer it through measurement.

$$F \xrightarrow{\text{measurement procedure}} R \xrightarrow{\text{interpretation}} X(F).$$

The intermediate step is often treated as background. In practice, though, the reading R is generated under a measurement condition C that encompasses the instrument specification, the training procedure, and the resulting trained instance. Describing the observation as $R(F)$ therefore obscures a potentially important element of the measurement process: it implicitly assumes that the reported quantity is insensitive to the relevant choices within C.

Most representation-analysis work varies F while holding the measurement procedure constant; we reverse the setup: we keep F fixed and intervene on C. This reframes what is often a tacit assumption as an empirical issue: when a quantity changes while the representation is held constant and only the measurement condition shifts, it cannot be credited to F alone unless one also appeals to an extra invariance argument.

1.2 THE INVARIANCE QUESTION

Prior work indicates that probe results depend on probe capacity, regularization, optimization, and the training setup, and that representation-similarity measures have their own reliability and validity

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.

limitations. More generally, measurement theory emphasizes that observed scores must be interpreted in light of the methods used to obtain them. Our focus here is more narrow and pragmatic:

> *When the representation is held fixed, is the measured magnitude of a representational difference invariant to the measurement condition?*

This connects to the question of whether a property can be recovered using different probes, or which probe should be considered more trustworthy. That a measurement can vary with the procedure used is not the novel point we are making; instead, we ask whether, for a fixed representational contrast, that procedural dependence can be set aside, i.e., whether the size of the measured effect carries over across controlled conditions.

To examine this, we hold the representation fed into the instrument bitwise identical while varying the measurement condition (see Section 3), then evaluate the same 45 contrasts in two trained conditions using an unchanged instrument specification. The pairwise structure stays strongly correlated even as the measured scale shifts markedly; normalizing by the operating point reduces, but does not fully remove, the discrepancy, and simpler probes recover the same qualitative behavior.

Crucially, we are not asserting that any intrinsic property of F varies with the measurement condition. The empirical claim is more pointed: even with F fixed, substantial relational agreement can coincide with a breakdown in magnitude transportability. Until such invariance is demonstrated, the appropriate interpretation is the condition-dependent quantity $X(F; C)$, not a property of F in isolation.

1.3 Contributions

- **Measurement-conditioned formulation.** We formulate representation analysis as $R = \mathcal{M}(F; C)$ and make invariance to the measurement condition an explicit object of study. The key question is not whether the measurement procedure exists, but whether its effect can be ignored for the reported quantity.
- **A controlled test of scale invariance.** We introduce a protocol that holds the measured representation bitwise fixed while changing the measurement condition. Applied to the same 45 representation pairs, two trained conditions yield strongly correlated pairwise effects but substantially different measured magnitudes. Operating-point normalization explains most, but not all, of the scale difference; the remaining pattern is consistent with an external $1 - \mathrm{CKA}$ reference and qualitatively recurs with simpler probe families.
- **What remains stable across conditions.** Although absolute magnitude is condition-sensitive, relative structure is substantially more stable: representation ordering and spatial agreement largely persist across effective trained conditions, while a degenerate saturated condition can make this structure effectively unobservable. These results distinguish *scale sensitivity* from *relational stability* rather than treating measurement dependence as an all-or-nothing failure.

2 Measurement-Conditioned Representation Analysis

2.1 Formalization

We write a measurement as

$$R = \mathcal{M}(F; C), \tag{1}$$

where F is the representation being measured, C is the measurement condition, and R is the resulting reading. Conventional notation often suppresses the measurement condition and writes the observation as a function of F alone; Equation 1 makes that dependence explicit.

We use $C \supseteq (I_0, E, W)$ to collect the components of the measurement procedure that can affect the reading, where I_0 denotes the instrument specification (e.g., architecture and hyperparameters), E its training environment (e.g., supervision signal, data, and optimization recipe), and $W = \mathrm{Train}(I_0, E, \xi)$ the resulting trained instrument instance under randomness ξ. These components need not be independent in parameter space; what matters experimentally is that the design specifies which are held fixed and which are changed (Appendix D).

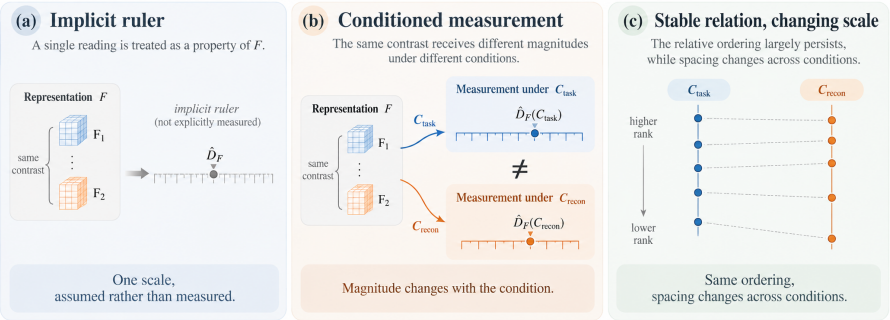

Figure 1: **Measurement-conditioned representation analysis. Left** (*implicit ruler*): conventional attribution suppresses the measurement condition and treats a measured representational difference as if it were expressed on a single fixed scale. **Middle** (*conditioned measurement*): for the same representational contrast, changing the measurement condition can change the measured magnitude, $D_F(C_1) \neq D_F(C_2)$. **Right** (*stable relation, changing scale*): relative structure can remain similar across effective conditions even when the measured scale changes. The experiments test these invariances rather than assume them.

Equation 1 is therefore a bookkeeping formalism rather than an empirical result; the scientific question is whether a quantity obtained through such a procedure is *invariant* across a specified family of measurement conditions. We test that question while holding the measured representation and observable fixed and changing the trained measurement condition.

2.2 Measured Effect Magnitude and Hypothesis

We focus on a concrete quantity: the measured magnitude of differences between representations under a measurement condition. For image k and representation pair (i, j), define the signed contrast $\delta_{ij,k}(C) = R_k(F_j, C) - R_k(F_i, C)$.

Our target quantity is the average absolute magnitude of these contrasts:

$$\boxed{D_F(C) \;=\; \mathbb{E}_{(i,j),k}\big[|\delta_{ij,k}(C)|\big]} \tag{2}$$

For a finite set of representation pairs $\mathcal{P}$ and n images, we estimate this quantity by the mean-of-absolute-differences convention $\widehat{D}_F(C) = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \frac{1}{n}\sum_{k=1}^{n}|\delta_{ij,k}(C)|$, used for the primary analysis; alternative aggregation conventions and their sensitivity analyses are reported in Appendix B.

The central invariance question is whether changing the measurement condition changes this magnitude. Formally, we test the null hypothesis $H_0^{\text{scale}} : D_F(C_1) = D_F(C_2)$. The experiments in Section 4 evaluate it using the empirical estimates $\widehat{D}_F(C)$; signed factorial contrasts and related decompositions address different quantities and are reported separately in Appendices B and F.

Figure 1 summarizes the analysis: we first ask whether changing C changes the reading of a fixed representation, then whether the measured magnitude of a fixed batch of representational contrasts changes across conditions, and finally which relational properties such as ordering and spatial agreement remain stable; the design in Section 3 implements these comparisons with the measured representation held fixed.

3 Experimental Design

3.1 Holding the Representation Fixed

The central requirement of our design is to vary the measurement condition without changing the representation being measured. We therefore freeze the backbone throughout measurement and attach the instrument through a zero-initialized residual interface $\text{out} = F + s \cdot \text{out_proj}(G)$. At

attachment time, all weights and biases of `out_proj` are zero, so out $= F$ holds exactly for any input, making the initial interface an algebraic identity rather than an approximation.

During instrument training, the backbone remains frozen, and BatchNorm statistics are kept fixed. We verify tensor by tensor that the representation F entering the instrument is bitwise identical across measurement conditions ($\max|\Delta F| = 0$; Appendix A). Interventions on the measurement condition are thereby experimentally separated from changes to the measured representation; implementation-level checks (BatchNorm verification, effect of the trained residual branch on the host output) are reported in Appendix A.

3.2 Measurement Conditions and Observable

The primary comparison uses two trained measurement conditions with the same instrument specification: C_{task}, trained with segmentation cross-entropy, and C_{recon}, trained with reconstruction MSE. They therefore share I_0 but differ in the training environment E and the resulting trained instance W. We additionally include an untrained condition, C_0, only as a boundary condition for saturation and loss of observability, not as part of the primary comparison; full configurations are in Table 1.

The main observable is state entropy: at each spatial location the instrument maps the representation to a distribution P over K states, and the resulting entropy is aggregated into a scalar reading H. Because $H \in [0, \ln K]$, the observable has an explicit ceiling determined by the instrument state count rather than by the number of task classes. This bounded scale makes the instrument's operating point explicit and motivates the scale-normalization analysis in Section 4.3. We use $K = 16$ in the main experiments; the hyperparameter scan is reported in Appendix E.

3.3 Experimental Panel and Statistical Inference

The main panel contains 10 host models, formed as a 5×2 design over finetuning recipes and seeds (five learning-rate/weight-decay settings, two seeds each; Appendix F lists them), yielding all $\binom{10}{2} = 45$ representation pairs. Because the recipe grid deliberately spans a range of finetuning configurations, this panel is *not* a set of mutually indistinguishable representations: we therefore read it as the full panel (10 hosts, 45 pairs), whose pairs carry the primary scale comparison, and use a strict cohort, the largest mutually indistinguishable subset under the criterion $\mathrm{CKA} \geq 0.989$ (Appendix F), for the claims that assert indistinguishability and for the CKA-threshold sensitivity analysis. We compute measurements on a fixed 256-image diagnostic subset of PASCAL VOC 2012 using a DeepLabV3-ResNet50 host, with the backbone frozen throughout measurement.

The 45 pairwise contrasts are not 45 independent samples: they share the same 10 hosts, so inference for the primary scale ratio resamples hosts as clusters, and image-level quantities are resampled at the image level. We use percentile bootstrap intervals throughout; full resampling conventions and parameters are given in Appendix B, and cross-backbone supporting evidence with ConvNeXt-T in Appendix E.

3.4 Measurement Instrument

Our primary apparatus, RouteScope, maps each spatial feature to a distribution over K states through local state propagation and reads out the resulting state entropy; full architectural and mathematical details are in Appendix A.

The interface is instrument-agnostic: RouteScope, a linear probe, and a one-hidden-layer MLP attach to the same frozen representation interface. RouteScope is therefore an experimental apparatus rather than the object of study, which is the measurement relation $R = M(F; C)$ and its dependence on C.

4 REPRESENTATION EFFECTS DEPEND ON THE MEASUREMENT CONDITION

4.1 FIX THE REPRESENTATION, VARY THE MEASUREMENT CONDITION

To isolate measurement dependence at the reading level, we hold the representation fixed and vary only the trained instrument instance. Under the path-B protocol, three independently initialized and trained instances receive bitwise-identical backbone representations ($\max|\Delta F| = 0$; Appendix A), yet produce clearly separated readings: $H = 1.1917$, 0.9975, and 0.7109. By contrast, before training, independently initialized instruments differ by only 0.00315 in H, compared with a range of 0.48 after training (Appendix F). Thus, even when the measured representation is unchanged, the absolute reading can depend strongly on the trained instance, motivating the sharper question below: does the measurement condition also change the measured magnitude of differences between representations?

4.2 SAME REPRESENTATIONAL CONTRASTS, DIFFERENT MEASURED SCALES

We now test the paper's central invariance question. The same 10 frozen representations define $\binom{10}{2} = 45$ representation pairs, and each pair is measured under two trained conditions with the same instrument specification: C_{task}, trained with segmentation supervision, and C_{recon}, trained with reconstruction supervision.

The same representational contrasts acquire substantially different measured magnitudes under the two conditions. The measured effect magnitude is

$$\widehat{D}_F(C_{\text{task}}) = 2.38 \times 10^{-2}, \qquad \widehat{D}_F(C_{\text{recon}}) = 4.02 \times 10^{-3},$$

giving a paired scale ratio of $5.93\times$ (95% CI $[5.71, 6.15]$; Table 8 in Appendix F). Thus, for these two trained measurement conditions, we reject the hypothesis of equal measured effect magnitudes. As a supplementary instance-sensitivity check, independently retraining five instrument instances per condition yields paired ratios in the same range (5.7–6.0; Appendix B), suggesting that the observed difference is not specific to a single trained instance.

Importantly, the change is predominantly one of *scale rather than pairwise structure*: across the same 45 representation pairs the effect magnitudes remain highly correlated ($r = 0.975$), with a scatter that is approximately proportional but displaced from the equal-scale reference (Figure 2a). Pairs that appear relatively more different under one condition therefore tend also to appear more different under the other, while the magnitude assigned to those differences changes systematically.

The numerical factor $5.93\times$ is descriptive of these conditions, not a universal conversion constant: the claim tested here is the failure of scale invariance ($\widehat{D}_F(C_{\text{task}}) \neq \widehat{D}_F(C_{\text{recon}})$), not the universality of the observed ratio.

4.3 ACCOUNTING FOR SCALE COMPRESSION AND INSTRUMENT SPECIFICITY

Because H is bounded ($H \in [0, \ln 16]$), the most immediate alternative explanation is that the $5.93\times$ gap arises simply because the two conditions operate at different positions on the entropy scale. Our question is not whether effect magnitudes are invariant to arbitrary numerical reparameterizations, since a rescaling of the observable necessarily rescales its differences, but whether the measured magnitude is transportable across experimentally realized measurement conditions. We therefore quantify how much of the observed difference is explained by the operating point, and test whether the remaining non-invariance is consistent across alternative references, instruments, and host subsets.

Operating-point normalization. C_{task} and C_{recon} operate at 61.8% and 90.9% of the entropy ceiling, so scale compression is a genuine concern. Available-range normalization explains most of the raw difference ($5.93 = 4.19 \times 1.42$), leaving a residual $1.42\times$ (95% CI $[1.36, 1.47]$; Figure 2b). The raw $5.93\times$ is therefore descriptive; the result is that equal-scale invariance is not restored after normalization (other monotone normalizations and a gain model: Appendix C).

External reference. As an external consistency check, we regress the pairwise effect magnitudes against $1-\text{CKA}$, which does not depend on the learned instruments, obtaining a scale ratio of $5.91\times$

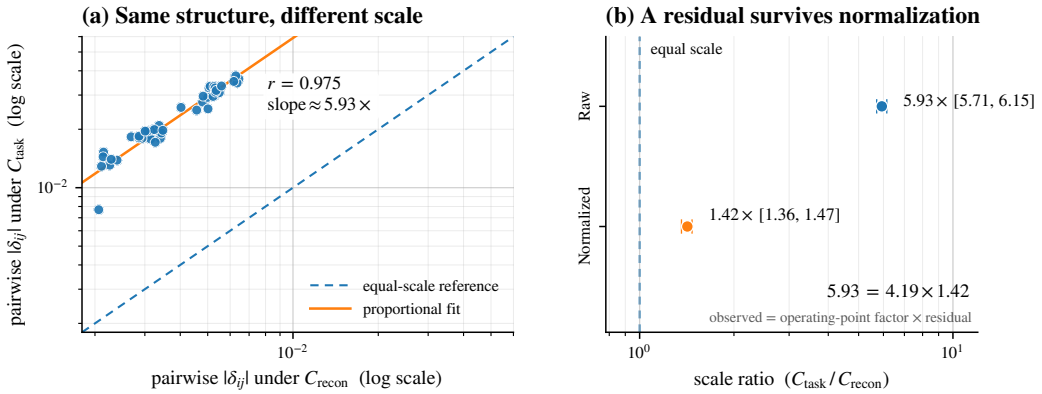

Figure 2: **Highly correlated pairwise structure, different measured scale. Left:** effect magnitudes for the same 45 pairs under C_{recon} and C_{task}: strongly correlated ($r = 0.975$) but displaced from the equal-scale reference, with a proportional fit corresponding to a $5.93\times$ scale difference. **Right:** operating-point normalization explains a factor of 4.19 of the raw difference, leaving a residual $1.42\times$ effect (95% CI $[1.36, 1.47]$), still above the equal-scale value of 1.

that closely matches the direct $5.93\times$ ratio. This is consistent with the two conditions preserving much of the same pairwise structure while assigning it different scales (regression diagnostics in Appendix C).

Alternative instruments. The failure of scale invariance is not specific to RouteScope. A linear probe with no learnable routing structure and a matched one-hidden-layer MLP both reproduce the qualitative interaction: under the same two training conditions, their ratio confidence intervals exclude 1 (Appendix D). Because the reconstruction-side probes operate near saturation, we treat these experiments as qualitative cross-instrument replications rather than estimates of a common numerical scale factor.

Host sensitivity. The result is not driven by a single representation: leave-one-host-out recomputation yields ratios in $[5.88, 6.03]$ across all ten exclusions (Appendix C).

Together, these checks show that operating point explains much, but does not restore equal-scale invariance; the remaining qualitative pattern is also consistent with the external reference and recurs in simpler probe families.

5 What Survives Across Measurement Conditions?

If absolute magnitude is condition-sensitive, what, if anything, remains stable across measurement conditions? We find that relative structure is substantially more stable than absolute scale, although it is not invariant. Figure 3 summarizes this pattern along three dimensions: the ordering of representations, their spatial agreement, and the conditions under which such structure becomes difficult to observe.

5.1 Relative Ordering Is More Stable than Absolute Scale

The scale change of Section 4.2 does not erase the relative organization of the representations: across all 10 hosts the two trained conditions yield significantly concordant rankings (Spearman $\rho_s = 0.7333$, Kendall $\tau = 0.6889$; exact permutation test $p = 0.0101$, $n = 10$), with 8 of the 10 hosts within one rank. The agreement is nonetheless not exact: seven of the 45 pairs reverse order, most on a single host whose training had degenerated, and the saturated C_0 condition yields no consistent ordering relative to either trained condition ($|\rho_s| < 0.18$; Appendix B). Relative ordering is therefore more stable across effective measurement conditions than absolute magnitude, but should not itself be treated as invariant.

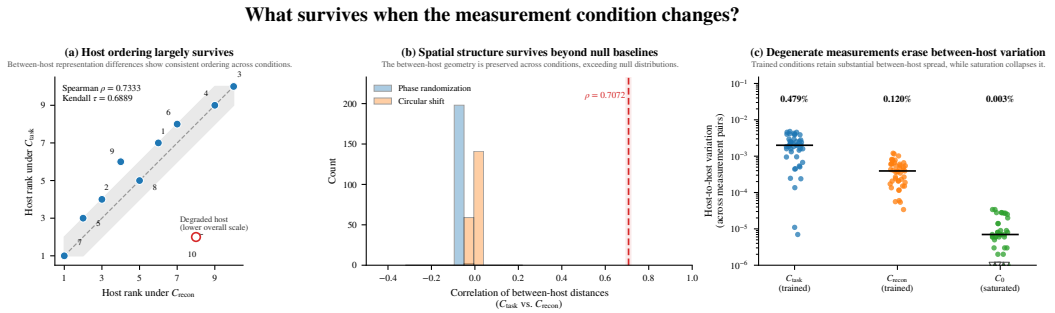

Figure 3: **Relational structure is more stable than absolute scale, but can become effectively unobservable under a degenerate measurement condition. Left:** host rankings under the two trained conditions are significantly concordant; 8 of 10 hosts move by at most one rank, while one degraded host moves six ($\rho_s = 0.7333$, $\tau = 0.6889$, permutation $p = 0.0101$). **Middle:** cross-representation spatial agreement exceeds all samples from both surrogate nulls (phase randomization and circular shift; 200 samples each), so it is not explained by these null structures. **Right:** host-to-host variation remains measurable under the two trained conditions (0.479% and 0.120% of the condition mean) but nearly collapses under the saturated C_0 condition (0.003%).

5.2 Spatial Agreement Persists Across Representations

The same pattern extends beyond scalar rankings to spatial organization. Under the main R_{entropy} convention with strict BatchNorm control, pooling three representation pairs over 256 images gives cross-representation spatial agreement $\rho = 0.7072$ (95% CI $[0.6958, 0.7176]$), which exceeds every sample from both surrogate null distributions (phase randomization and circular shift; Appendix E), so the correspondence is not explained by these nulls. A per-pixel instrument with no spatial propagation yields even stronger agreement on the same representations and supervision ($\rho_{\text{entropy}} = 0.9912$), showing that high cross-representation spatial agreement does not require RouteScope's routing or propagation mechanism; the lower agreement under the structured instrument is therefore consistent with the measurement procedure modifying the strength of the observed correspondence rather than creating it (Appendix E).

5.3 Degenerate Conditions Can Obscure Relational Structure

The preceding results also expose a boundary condition: measurement conditions can affect not only the magnitude of an observed effect, but also whether relational structure remains observable under the chosen measurement. Under the untrained C_0 condition, the entropy reading is nearly saturated, host-to-host variation collapses to 0.003% of the condition mean, and no consistent ordering can be recovered. This loss of observability should not be equated with saturation alone: C_{recon} operates at 90.9% of the entropy ceiling yet still preserves substantial relational ordering, whereas the near-collapse occurs for C_0 at 99.6%. In our experiments, relational structure remains visible under the trained, non-degenerate conditions but becomes effectively unobservable under the saturated boundary condition.

> Measurement conditions can change the measured scale,
> and can affect whether relational structure remains observable. (3)

6 Related Work

Representation similarity. A large literature compares neural representations using RSA, SVCCA, PWCCA, CKA, and related measures (Kriegeskorte et al., 2008; Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019), examines the reliability, statistical behavior, and failure modes of such similarity measures (Ding et al., 2021; Davari et al., 2023; Klabunde et al., 2025), or asks through model stitching whether representations can be functionally substituted across networks (Lenc &

Vedaldi, 2015; Bansal et al., 2021; Raghu et al., 2021). These methods compare representations under a chosen metric or functional test; our question is complementary: for a fixed representational contrast, does the *magnitude attributed to it* remain invariant when the measurement condition changes?

Probing. Probe-based analysis asks what information can be decoded from a representation and how that conclusion depends on the probe (Alain & Bengio, 2017; Belinkov, 2022; Hewitt & Liang, 2019; Pimentel et al., 2020; Voita & Titov, 2020; Ravichander et al., 2021). Prior work shows that probe capacity, regularization, optimization, control tasks, and training recipes can substantially affect probe-based conclusions. Our setting overlaps with this concern but asks a different estimand: rather than whether a property is decodable, or which probe is preferable, we hold the representational contrast fixed and test whether the *measured effect magnitude* assigned to it is transportable across controlled measurement conditions; probe dependence is thus one instance of the broader measurement-conditioning problem studied here.

Measurement theory. Generalizability theory, measurement invariance, and construct validity (Cronbach et al., 1972; Meredith, 1993; Cronbach & Meehl, 1955; Messick, 1995; Borsboom et al., 2004) establish that an observed score should be interpreted relative to the procedure that produces it, and related ideas have increasingly informed machine-learning evaluation (Blodgett et al., 2020; Jacobs & Wallach, 2021; Liao & Xiao, 2023). This tradition motivates our framing; we do not propose a new theory of measurement invariance. Our empirical contribution is to operationalize its invariance question for neural representation analysis and show, under a controlled fixed-F intervention, that strong relational agreement can coexist with failed magnitude transportability.

Representation interventions and under-specification. A related line of work uses causal interventions, activation manipulations, or model editing to study the role of internal representations and computations (Geiger et al., 2021; 2024; Vig et al., 2020; Zhang & Nanda, 2024; Ravfogel et al., 2020; Meng et al., 2022; 2023). Those methods intervene on the representation or on the computation that produces behavior, whereas our intervention is orthogonal: we keep the measured representation fixed and intervene on the measurement procedure applied to it. Under-specification further shows that seemingly equivalent training specifications can yield different model instances (D'Amour et al., 2022); in our formulation the trained measurement instance W is therefore part of the measurement condition rather than an implementation detail.

Taken together, these lines of work motivate the distinction at the center of this paper: representation analysis should separate *variation in the representation* from *variation induced by the procedure used to measure it*. Our contribution is a controlled test of whether a reported representational effect is invariant across such measurement conditions.

7 Discussion and Limitations

7.1 Conditioned Representational Claims

The central implication of our results is that a representational quantity obtained through a measurement procedure should be interpreted relative to the conditions under which it was measured. We denote this explicitly as a *measurement-conditioned claim*, $X(F; C)$, rather than writing the quantity as if it depended on F alone.

> **Measurement-conditioned claim.** A measured representational quantity may be attributed to F alone only to the extent that it is invariant across the measurement conditions over which the claim is intended to generalize; otherwise, the measurement condition is part of the scope of the claim.

This distinction matters for magnitude-based claims—for example, whether two representations are "nearly identical," whether one intervention has a larger representational effect than another, or whether effect sizes are comparable across studies—even when rankings agree. In practice, a minimal audit is to report the operating point and uncertainty, repeat the measurement across plausible conditions or trained instances, and report relational agreement separately from magnitude agreement. If the magnitude conclusion changes materially, the claim should remain conditioned on C; Table 5 in Appendix B gives the reporting set.

7.2 Relation to Representation Similarity and Probing

Measurement-conditioned analysis complements representation-similarity methods and probing rather than replacing them (Section 6). Similarity methods characterize relations between representations under a chosen metric, while probing asks what information a chosen probe can extract. Our question concerns another axis: whether a representational effect retains its measured magnitude when the measurement condition changes.

The main experiment illustrates why these views can coexist: pairwise effect magnitudes remain strongly correlated across the two trained conditions ($r = 0.975$) even though their absolute scale differs substantially. Relational agreement therefore does not imply scale invariance, while failure of scale invariance does not imply an absence of stable representational structure. This distinction between *relational stability* and *magnitude transportability* is what the measurement-conditioned formulation makes explicit.

7.3 Limitations

L1 (scope across architectures and modalities). The empirical study is limited to vision representations, with DeepLabV3-ResNet50 as the main host and ConvNeXt-T providing cross-backbone supporting evidence. We therefore make no claim that the observed magnitude effects, operating-point behavior, or residual normalization factors transfer quantitatively to language or multimodal representations. The formalism $R = \mathcal{M}(F; C)$ is modality-agnostic, but the relevant conditions and observables may differ across domains.

L2 (measurement-instance variability). The primary quantitative analysis is reported for one trained instrument instance per condition. A supplementary check with five independently retrained instances per condition yields similar paired ratios (Section 4.2, Appendix B), reducing the concern that the result is specific to a single checkpoint. This check remains within the same instrument family, training recipe, and host panel, so we do not claim a population-level scale factor across arbitrary trained instruments. The broader claim is failure of invariance under the conditions studied here, not universality of the observed ratio.

L3 (construct and measurement scale). State entropy is an operational measurement scale, not an intrinsic semantic property of the representation. Effect magnitude therefore depends on the chosen observable and its calibration; we report cross-observable and normalization sensitivity analyses in Appendix C. Our conclusions concern failure of invariance for the measurements studied here, not identification of a "correct" ruler, a universal entropy-based representational distance, or a direct semantic readout for individual predictions (Appendix E).

8 Conclusion

Representation analysis does not observe a representation in isolation but through a measurement procedure. We make this dependence explicit as $R = \mathcal{M}(F; C)$ and test whether the magnitude assigned to a fixed representational contrast is invariant to the measurement condition.

Holding the measured representations bitwise fixed, we find that the same contrasts can preserve highly similar relational structure while receiving different measured magnitudes. Operating-point normalization explains most, but not all, of the scale difference, and simpler probe families reproduce the qualitative failure of equal-scale invariance. Relative ordering and spatial agreement are substantially more stable than absolute magnitude, though not invariant, and can become effectively unobservable under a degenerate condition.

These results motivate a simple reporting principle: a measured quantity should be attributed to F alone only over conditions for which the relevant invariance has been established; otherwise the condition is part of the claim's scope, written explicitly as $X(F; C)$.

More broadly, invariance to the measurement procedure should be treated as an empirical property of a representational claim rather than an assumption built into it.

AI USE STATEMENT

We used generative AI tools to assist with language editing, presentation and figure-design discussions, visualization-code drafting and review, and bibliographic verification. All AI-assisted text, code, figures, citations, and technical claims were reviewed and verified by the authors. Generative AI was not used to generate experimental data or to replace the execution and verification of the reported experiments. We take responsibility for the final content of this work, including all material produced with the aid of generative AI.

ETHICS STATEMENT

This paper studies the measurement methodology of representation analysis. All experiments use public datasets and public models, and involve no human subjects, personally identifying information, or sensitive data. Property claims in representation analysis are often used to support downstream decisions, so the auditability of such claims has fairness implications (limitation L3 of Section 7.3). We argue that auditability should be increased: when the measurement condition is explicitly registered, the scope of a claim such as "F has X" becomes explicit, reducing the risk of mistaking a measurement artifact for a property of the representation. This paper proposes no new capability or application system, and we identify no foreseeable misuse.

REPRODUCIBILITY STATEMENT

Data and protocol. The main experiment is PASCAL VOC 2012 semantic segmentation with a DeepLabV3-ResNet50 host; cross-backbone supporting evidence uses ConvNeXt-T (Appendix E). All diagnostic readings come from a fixed 256-image validation subset (seed $= 42$); the backbone is frozen during measurement, and BatchNorm is canonicalized along path B (Appendix A). Fixing F is guaranteed by the algebraic-identity bypass, whose bitwise agreement is verified with `torch.equal` and $\max|\Delta F| = 0$, and can be recomputed by script; Appendix A (Tables 2 and 3) reports the limited effect of the interface's additive term on host behavior and its source. Each of the three classes of controlled intervention comes with a complete configuration, random seed, and checkpoint identifier (Appendices A and D).

Statistics. All intervals in this paper are cluster bootstrap (percentile, seed $= 42$; the resampling unit and B are declared with each convention, $B = 2000$ or 4000). The paired ratio of the main result uses host-level resampling (the resampling unit is the 10 hosts, $B = 4000$); all remaining contrasts are image-level. Reported quantities are means over repeat units (random seeds, hosts), and the resampling parameters and convention name of every contrast are registered alongside the number in the convention index of Appendix B.

Code and provenance. Model, intervention protocol, statistics, and plotting code are released with the paper; training checkpoints are too large to include and can be retrained from the given configurations and seeds. Provenance rules (the ledger and the generating script behind every load-bearing number, the drift checks guarded by `assert`, and the treatment of historical readings) are given in Appendix G.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2017.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. doi: 10.18653/v1/2020.acl-main.485.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. The concept of validity. *Psychological Review*, 111(4):1061–1071, 2004.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 2022. arXiv:2011.03395, 2020.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity with statistical testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/D19-1275.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 375–385, 2021.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3728458.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. doi: 10.1109/CVPR.2015.7298701.

Q. Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993. doi: 10.1007/BF02294825.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9):741–749, 1995.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. doi: 10.18653/v1/2020.acl-main.420.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. doi: 10.18653/v1/2020.acl-main.647.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.14.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

A Experimental Protocol and Implementation

A.1 Complete mathematical details of SGLNDP and instrument configuration. The transition kernel of the SGLNDPBlock, $T = \mathrm{RowNorm}(A \odot S \odot E)$, is built from three factors: a positional amplitude gate A (computed from the input features, gating the strength of neighborhood messages, with a tunable sharpness α), a directional softmax S (allocating messages across directions), and a channel modulation gate E (modulating per channel, with tunable sharpness β, γ). Each layer performs a discrete diffusion update as a residual increment: $Z^{(t+1)} = Z^{(t)} + \lambda\,(TZ^{(t)} - Z^{(t)})$. Reflecting boundaries handle image edges: neighborhood positions beyond the boundary are mirrored, preserving message conservation at the boundary. The Markov constraint is enforced by row normalization, so each row of T sums to one. The scan over the hyperparameters α, β, γ is in Appendix E. This module is an analytical tool within the measurement instrument; we do not claim that the backbone performs the same operations internally. The reading H is invariant to permutations of the state numbering and can be compared directly across instruments; the K components of P are not compared across instruments (state labels are defined by the projection and are not aligned after independent training).

Reading path. The readings of this paper are taken from the per-pixel routing distribution P returned by `\_core\_forward`; `out\_proj` participates only in the additive bypass returning to the host, not in the reading. Hence the post-training host-behavior drift discussed below affects only the host output, not R.

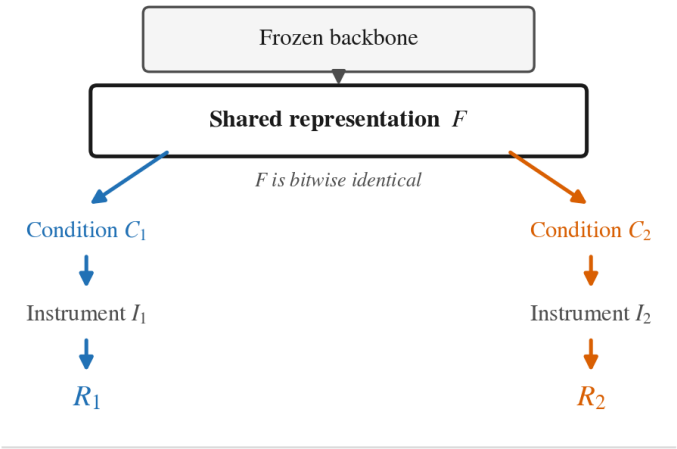

Figure 4: The non-destructive measurement interface (device diagram, Appendix A). A frozen backbone branches, through the frozen representation F, into two measurement arms $C_1 \rightarrow$ instrument $\rightarrow$ reading R_1 (blue) and $C_2 \rightarrow R_2$ (orange). The two arms share the same bitwise-identical F, and the intervention acts only on the measurement condition C: *change the measurement condition, not the representation*.

Table 1: Complete configuration of the three measurement conditions.

Item	Value
Architecture	the same `E8Plugin`: $c_{\text{in}} = 2048$, $c_{\text{state}} = 64$, $K = 16$, $l = 3$, `out\_mode=residual`, `multi\_collapse=True`
Parameter count	5.95×10^6 (identical for all three)
Training recipe	SGD (momentum 0.9), lr 10^{-2}, weight decay 10^{-4}, cosine annealing, 20 epochs, batch 32, input 321×321, AMP
Frozen scope	`--freeze-base`: backbone, segmentation head and the $768 \rightarrow 2048$ projection are all frozen; only instrument parameters are trainable; BatchNorm follows path B (`--eval-backbone`, `backbone.eval()` at every step)
E1	objective = VOC segmentation CE; `residual\_scale\_init=1.0`; checkpoint `runs/20\_PLUGIN\_TASKS/voc\_seg\_e8/e8\_curves/C1C4\_seed123/best\_model.pt`
E2	objective = reconstruction MSE; `residual\_scale\_init=1.0`; checkpoint `runs/20\_PLUGIN\_TASKS/voc\_seg\_e8/pretrained\_recon\_freeze\_ep20/best\_model.pt`
E3	untrained (no checkpoint); `residual\_scale\_init=0.0`; initialization seeds 777/778/779, per-image readings averaged over the three seeds

A.2 Complete configuration of the three measurement conditions. $C_{\text{task}}/C_{\text{recon}}/C_0$ of the main text correspond to E1/E2/E3 of Table 1. The three are architecturally identical and differ only in the training objective, checkpoint, and residual-bypass initialization; the instrument specification I_0 is fixed, and we compare the training environment E and the instance W.

A.3 Identity verification and post-training drift. The zero-initialized residual bypass holds exactly when an untrained plugin is attached: VOC DeepLabV3-R50 has mIoU 71.07 both before and after attachment ($\Delta = 0$). Tensor-by-tensor comparison covers all 182 backbone weight tensors, which are exactly equal. After training, a limited drift appears, $\Delta = -0.06$ (71.01). Its mechanism is checkable and contains neither gradient leakage nor a BatchNorm interaction:

This non-destructiveness is guaranteed by method design, not by experimental induction.

Table 2: Candidate mechanisms for the post-training drift, and their verification.

Candidate mechanism	Holds	How it was checked
Gradients flowing back into the backbone	×	under `--freeze-base`, backbone, segmentation head and projection all have `requires\_grad\_(False)`
Drift of BatchNorm statistics	×	path B calls `backbone.eval()` at every step; A.4 verifies layer by layer that the drift of all 53 BN layers is 0
Non-zero additive interface term after training	✓	out $= F + s \cdot$ out_proj(G); `out\_proj` is zero-initialized and no longer zero after training, so the segmentation head sees out $\neq F$

Table 3: Item-by-item check before and after attachment.

Check	Before attachment	After attachment (untrained instrument)	After training
Backbone representation F bitwise identical		$\max\|\Delta F\| = 0$, `torch.equal=True`	backbone frozen, still 0
CKA with the pre-attachment representation		1.0000	1.0000
VOC mIoU	71.07	71.07 ($\Delta = 0$)	71.01 ($\Delta = -0.06$)

A.4 BatchNorm canonicalization (path B). Path B keeps `model.train();model.backbone.eval();model.classifier.eval()` in the training loop, so that the frozen backbone's BN statistics do not drift with the seed. The three models M1/M2/M3 produce bitwise identical forward outputs F during the measurement phase: $\max|\Delta F| = 0$, `torch.equal=True`, CKA $= 1.0000$; the `running\_mean`/`running\_var` drift of all 53 BN layers is 0.

B Statistical Estimands and Resampling

B.1 Definition of the main estimand. The main effect magnitude throughout the paper is mean-of-abs (estimator B), that is, the scale estimand $\widehat{D}_F(C)$ of Section 2.2:

$$\boxed{\widehat{D}_F(C) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{1}{n} \sum_{k=1}^{n} \big| R_k(F_i, C) - R_k(F_j, C) \big|} \tag{4}$$

where $\mathcal{P}$ is the set of representation pairs ($\binom{10}{2} = 45$ for the main panel) and n the number of per-image samples ($n = 256$, seed $= 42$). All other estimators (abs-of-mean and the rest) serve only as sensitivity analyses, with conventions in Table 4. Estimator A (abs-of-mean) returns a different number for the same comparisons: on this paper's main 45-pair panel it is $3.32\times$ (host-level paired bootstrap CI $[2.57, 4.74]$), while the $4.09\times$ registered for this convention is its value on the 10-pair strict cohort (F.4), not on the main panel (two panels under one convention, and neither may be used interchangeably with the $5.93\times$ of the main convention). The numerical ratio therefore depends on the estimand and the panel; rejecting equal scale does not, since both host-level intervals lie far from 1.

B.2 Table 4: the convention index. Every convention name used in the main text and in the tables is defined here. In this project, almost all numerical disputes have come from "the same name under different conventions", so this table is the entry point for reading our numbers.

B.3 The 2×2 factorization (signed supplementary analysis, not the same object as the main estimand). Let $R_{ij} = R(F_i, C_j)$. The representation main effect is $\Delta_F = \frac{1}{2}[R_{21} + R_{22} - R_{11} - R_{12}]$, the measurement-condition main effect is $\Delta_C = \frac{1}{2}[R_{12} + R_{22} - R_{11} - R_{21}]$, and the interaction is $\Delta_{FC} = R_{22} - R_{21} - R_{12} + R_{11} = \delta_{ij}(C_2) - \delta_{ij}(C_1)$. The interaction acts on the **signed** pairwise contrast $\delta_{ij}(C)$, whereas the main estimand $D_F(C)$ takes the absolute value first: the two are not the same object, and $\Delta_{FC} = 0$ and $D_F(C_1) = D_F(C_2)$ do not imply one another. The primary claim of this paper tests $H_0^{\text{scale}} : D_F(C_1) = D_F(C_2)$; the signed 2×2 analysis is registered as supplementary evidence in F.6.

Table 4: The convention index.

Convention name	Definition	Appears in	Comparable across tables
Estimator A (abs-of-mean)	per-pair differences averaged over images first → absolute value → averaged over pairs	the two-arm formula of B.3	× not interchangeable with B
Estimator B (mean-of-abs)	per-image $\lvert\delta\rvert$ → averaged over images and pairs	the paper's main effect magnitude, Table 8, the $5.93\times$ of Section 4.2	× as above
Single-arm / two-arm	two-arm = the 2×2 factorial formula; single-arm = fix the instrument and measure the effect magnitude, fix the host and measure Δ_C	the main estimand $D_F(C)$ of Section 2.2 uses the single-arm convention; Δ_{FC} uses the two-arm convention (signed supplementary analysis, B.3 and F.6)	Δ_C is bitwise identical under both; Δ_F/Δ_{FC} are signed supplementary analyses and not the same object as D_F
Flagship pair / cohort-pooled / 10-pair aggregate	the resampling unit differs (1 pair / 5 hosts pooled / 10 pairs)	declared wherever used in the main text	the ratio is insensitive to how pairs are taken; the magnitude of $\widehat{D}_F(C)$ is not comparable
Measurement panel	path B (retrained with `\_eval`, backbone frozen, BN not drifting) / community-standard panel (original checkpoints, BN updated during training)	Section 4.1, F.1, F.5	× readings cannot be carried across panels: path B 1.1917 vs. community-standard panel 1.7136 (a difference of 0.522, containing both weight and host-panel differences)
Operating point / dynamic-range headroom	the distance of the reading baseline H from the scale ceiling (headroom $= \ln 16 - H$)	Sections 3.2 and 4.3	× the operating point is itself an instrument property and is not compared across instruments; it is used to normalize effect magnitudes
External reference ruler calibration	through-origin regression of the pairwise $\lvert\delta_{ij}\rvert$ on $1 - \mathrm{CKA}$, taking the slope	Section 4.3 and C.3	✓the ratio of slopes is comparable across instruments; a single slope carries that instrument's scale and is not
Transform-family normalization	after applying a monotone transform g to the observable, recompute ρ_g and $\Phi_g = m_a^g/m_b^g$, then take $r_g = \rho_g/\Phi_g$	C.2	✓r_g is the robustness quantity comparable across transforms; ρ_g is not
Instance-sensitivity check	five independently trained instances per condition (same instrument specification and training recipe, ξ = initialization seed and data order)	Section 4.2, B.7	× descriptive only: the same host-level paired ratio is recomputed per instance and clusters in a narrow range; no population-level interval across instances is claimed

B.4 Table 5: the minimal reporting set for $X(F;C)$. The minimal set that should accompany a property claim (Section 7.1 lists only the outline):

B.5 Relation to variance shares in generalizability theory. In a balanced design, the between-group sum of squares of a two-level contrast is proportional to the square of the contrast magnitude ($SS_A = (N/4)\Delta_A^2$, with N the total sample size), and the proportionality constant is the same for every contrast used here, so a ratio of effect magnitudes equals the square root of the corresponding

Table 5: The minimal reporting set for $X(F;C)$.

Item	Content
F	the identity of the measured representation (host, layer, checkpoint, diagnostic subset)
I	the instrument specification and instance (architecture, parameter count, training objective, random seed)
Measurement condition C	the values of (I_0, E, W); which components were fixed and which intervened on
Operating point	where the reading sits on the scale, and its dynamic-range headroom
Effect magnitude and interval	the value of $\widehat{D}_F(C)$ and its resampling interval (with resampling unit and B)
External reference	a consistency check independent of the learned instruments
What carries across conditions	shape and ordering conclusions, with their scope and known counterexamples

ratio of variance shares; the effect-size contrasts of this paper are convention-compatible with a variance decomposition. That relation only states convention compatibility and introduces no new number.

B.6 Base rates of the ordering reversals. Of the 45 pairs in the full panel, 7 pairs reverse sign between the trained instruments (E1–E2) ($7/45 \approx 15.6\%$), all of them inside the 17 pairs involving one training recipe (R5) ($7/17$), and 6 of those involve one and the same host, R5_seed310 (VOC mIoU 71.64, 4.42 below the lowest of the other nine hosts, rank difference 6). Removing it brings the full-panel reversal count to $1/36$ and Spearman to 0.9833; within the 10 pairs of the strict cohort, the reversal count is $0/10$. The main convention takes the full panel. Ledgers: `experiments/0Z_refactor_round/AE33\_hero\_fullpanel/`, `AE36\_number\_audit/`, `AE41\_pair\_cluster/`, `AE43\_transform\_family/`.

B.7 Instance-sensitivity check of the main ratio (Section 4.2). The main contrast is measured on one trained instrument instance per condition, and the host-level interval of the main convention conditions on those two instances. As a supplementary check of that dependence, five instrument instances were trained independently per condition (same instrument specification, training recipe and host panel; ξ = initialization seed and data order, seeds 401–405, 20 training epochs), and the paired ratio of the two conditions on the same 10-host panel was recomputed per instance: 5.838, 5.822, 5.933, 5.985, 5.705 (mean 5.857, instance sd 0.108), against the registered 5.928 of the main instance. The spread across instances is small relative to the distance of these ratios from 1, so the observed failure of equal-scale invariance is not specific to the main trained instance. The check is reported descriptively: it concerns the sensitivity of one quantity to the training instance W, not the distribution of the ratio over the instrument family, and it claims no population-level interval across instances. Ledger: `experiments/0Z\_refactor\_round/AE49\_multi\_instance/ae49\_multi\_instance.json`.

C Scale and Operating-Point Robustness

C.1 Operating-point normalization (Section 4.3). The three instruments have reading baselines $H = 1.7146/2.5198/2.7608$ (E1/E2/E3), which are $61.8\%/90.9\%/99.6\%$ of the ceiling $\ln 16 = 2.7725887$; the dynamic-range headroom values are $1.0580/0.2528/0.0118$, and the effect magnitude per unit headroom is $0.0225/0.0159/0.0176$ (the three differ by $1.42\times$). After headroom normalization the E1/E2 ratio is $1.417\times$, CI $[1.364, 1.470]$, and the identity $5.928 = 4.185 \times 1.417$ holds bitwise. The normalized residual for E1/E3 is $1.278\times$, CI $[1.243, 1.332]$.

C.2 Transform family and gain model (sensitivity check for Section 4.3). The transform family is $\{\text{identity},\ \exp(-H),\ \text{logit}(H/\ln K),\ m^{\beta}\ (\beta = 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0)\}$, ten in all; we write $\rho_g = \mathrm{E}|\Delta g|_a/\mathrm{E}|\Delta g|_b$, $\Phi_g = m_a^g/m_b^g$ and $r_g = \rho_g/\Phi_g$.

The AE35 five-task panel is the only other panel containing both E1 and E2, so it carries the transform-family check for the main conclusion; its host-level resampling has only 5 units, giving wide intervals (containing 1), whereas the image-level cluster bootstrap (resampling unit: images

Table 6: Transform-family normalization and the gain model.

Panel	Raw ratio ρ_g across transforms	Normalized residual r_g across transforms	Residual interval contains 1
AE35 five-task panel (E1/E2, includes E1)	1.410–78.679 (55.8×)	1.109–1.569 (1.41×)	image-level CI contains 1: 0/10; host-level ($n = 5$) CI contains 1: 10/10 (too few hosts in the panel)
AE23 R50 strict cohort (E2/E3)	0.570–4305.1 (7551×)	0.414–0.583 (1.41×)	0/10
AE26 six-host panel A/B (E2/E3)	2.483–497.3	2.047–2.532	0/10
AE42 linear probe ($C_{\text{task}}/C_{\text{recon}}$)	0.275–1.19×10^5	0.043–0.973	1/10 (see below)
AE42 one-hidden-layer MLP (same)	0.941–5.39×10^6	0.170–2.198	0/10

within a task) excludes 1, and both are registered alongside the number. **The single counterexample**: the AE42 linear probe with the logit transform gives $r = 0.973$, CI $[0.871, 1.157]$, containing 1; that panel's C_{recon} reading has $H \approx 2.754$, at 99.3% of the ceiling, a near-saturated degenerate instance whose denominator is already close to the precision floor.

The gain (compression) model. At 12 real operating points ($m \in [0.0541, 2.1487]$, covering both operating points of the primary scale comparison) we apply a small perturbation εu to the state distribution and measure the gain $|\Delta H|/\varepsilon$ ($\varepsilon = 0.05, 0.02$). Over 100 random directions plus the global radial direction, the model predicts an E1/E2 gain ratio in $[1.000, 3.500]$ (median 1.575; global radial 1.722; near-uniform analytic 2.046; architecture-specific non-parametric interpolation (the only architecture whose two operating points both lie inside the measured range) 2.036). The measured ratio 5.928 and the operating-point factor 4.185 are both above the upper end of that interval. Residuals after per-direction correction by the gain model: minimum 1.694, median 3.764, p_{95} 5.751, maximum 5.928 (global radial 3.442; near-uniform analytic 2.898). Fit quality is reported as it stands: the median R^2 of the per-direction log–log fits is 0.299, the maximum 0.884, and none exceeds 0.9, so the model should be read as a family of defensible normalizers, and the conclusion depends only on its upper bound.

C.3 External reference ruler (Section 4.3). On $1 - \text{CKA}$ the through-origin slopes are $3.503/0.593/0.031$ with $R^2 = 0.947/0.949/0.969$; the slope ratio $\text{E1}/\text{E2} = 5.907$ against a raw ratio of means of 5.928 (a difference of $\leq 0.8\%$). The range of $1 - \text{CKA}$ is about 10^{-3}, so the through-origin slope rests mainly on a few high-leverage points. **Three supplementary diagnostics (AE45, ledger-only, no retraining)** on the 45 pairwise values ($1 - \text{CKA}$ expanded from the 10×10 GAP-CKA matrix, ranging over $[2.06 \times 10^{-3}, 4.55 \times 10^{-2}]$; $\widehat{D}_F(C)$ as in AE33): (i) Regression with intercept: the slopes of all three instruments remain highly significant ($p < 10^{-9}$), the E1/E2 slope ratio with intercept is 5.80, consistent with the through-origin slope ratio 5.91 and the ratio of means 5.93; the intercept is significantly non-zero (E1 $\beta_0 = 1.16 \times 10^{-2}$), but this is an extrapolated value outside the data range (the smallest x is 2×10^{-3}), reflecting the non-linearity of the instrument response in the low-difference region, consistent with the operating-point/saturation account and not weakening "one shape, different scales". (ii) Pairwise rank correlation: Spearman $\rho = 0.917/0.920/0.914$ (E1/E2/E3, $p < 10^{-9}$), robust to the high-leverage points of the through-origin fit. (iii) Leave-one-host-out regression slope: the E1 slope falls in $[0.564, 0.603]$, E2 in $[0.097, 0.106]$, E3 in $[0.0054, 0.0057]$, not resting on any single host. Ledger: `experiments/0Z\_refactor\_round/AE45\_external\_cka\_diagnostics/ae45\_external\_cka.json`. **Convention note**: the $1 - \text{CKA}$ of this diagnostic uses the GAP-CKA matrix (a different pooling from the pairwise CKA of the through-origin slopes above), so the absolute slopes are not comparable across ledgers; the slope ratio of the two agrees (5.907, the same as the registered value), and the qualitative conclusions on intercept, rank, and leave-one-host-out are unaffected by the pooling; the two conventions differ only in the scale of the absolute slope, and that scale carries the convention of its own ledger and cannot be compared across ledgers.

C.4 Host sensitivity of the main ratio (Section 4.3, ledger-only, no re-running of inference). A leave-one-host-out over the 45 pairwise effect magnitudes (AE33 ledger), removing one host at a

time (36 pairs remaining each time) and recomputing the E1/E2 ratio of means, puts the ten ratios in $[5.88, 6.03]$, all far from 1; the per-pair ratios range over $[3.74, 7.11]$ with median 5.97. The generating script `scripts/refactor\_ae44\_hero\_scatter.py` guards the means ($2.3805 \times 10^{-2}/4.0157 \times 10^{-3}/5.928$), the correlation ($r = 0.9749$) and the reversal count ($7/45$) with `assert`. Ledger: `experiments/0Z\_refactor\_round/AE44\_leaveoneout/ae44\_leaveoneout.json` (source data `AE33\_hero\_fullpanel/`).

C.5 Cross-observable confirmation (comparability of H across instruments). The same pair of instruments gives different ratios on different observables: on ConvNeXt-T the E2→E3 ratio is $1.01\times$ by argmax diversity, $7.37\times$ by gap, $13.05\times$ by P_{primary}, and $24.69\times$ by H. All four scalar directions exceed 1, so the H reading does not depend on a single observable; at the same time, the ratios differ in magnitude, showing that no single conversion constant applies across measurement conditions: the instruments are not different markings on one ruler. This cross-confirmation is carried out only inside one instrument family and forms part of L2 of Section 7.3.

D Alternative Instruments and Environment Interventions

D.1 Non-orthogonality of the components of C, and environment interventions. E is not an independent factor that can be combined freely with I_0, because $W = \text{Train}(I_0, E, \xi)$ is a product of (I_0, E, ξ): changing E changes W, so "fix I_0 and move only E" is literally impossible; what can be fixed is I_0 and ξ. We therefore write the measurement condition as a bundle $C \supseteq (I_0, E, W)$ rather than as three independent factors: the components of C are intrinsically coupled, and what is needed is not orthogonality but that each component can be intervened on under control. The controlled environment intervention (AE32) fixes the instrument specification and the initialization seed and replaces the training environment as one package, so the objective-dependent routing and residual constraints change together with E (on the MSE side the residual bypass is zeroed, forcing the reconstruction to flow through the routing collapse; on the CE side the bypass is free). It therefore identifies a change-of-measurement condition attributable to the training regime, not an isolated loss-function effect. The intervention runs on three paired instrument instances (same frozen host, same initialization seed, same data order; only E swapped), so it exposes the instance component W rather than holding it fixed. The three instances give $\Delta_E \equiv H(E_a) - H(E_b) = +2.2329, +1.0966, +1.6316$ (mean 1.6537, instance sd 0.5685, a range of $2.04\times$): the sign is consistent across all three, but the instance spread is not negligible at the level of this contrast. The registered interval $\Delta_E = 1.6537$, CI $[1.6178, 1.6894]$, resamples the 768 pooled per-image values ($n = 3 \times 256$) and therefore does not cover the three instances; resampling the instances instead gives $[1.0966, 2.2329]$ (percentile bootstrap) and, with a t interval, $[0.2415, 3.0660]$ ($n = 3$). The within-recipe spread (8.188×10^{-4}) is a different and much smaller quantity and is not the uncertainty of this contrast. With three instances, this panel establishes the *direction* of the environment effect, not its magnitude. The training environment is not a disturbance that can be neglected but an identifiable component of the measurement condition; how much of the *magnitude* of Section 4.2 is attributable to W is addressed only descriptively by the instance-sensitivity check of L2 (Section 7.3, Appendix B). Ledger: `experiments/0Z\_refactor\_round/AE32\_ie\_split/ae32\_ie\_split\_result.json`.

D.2 The five probe arms. Five probe arms: D1 linear (66K parameters), D2 one-hidden-layer MLP (66K parameters), D3 RouteScope ($K = 16$, 5.95M parameters), D3s a $K = 1$ capacity-degenerate arm (3.92M parameters, of the same order as D3, separating routing capacity from parameter count), and D4 an untrained baseline. VOC segmentation CE supervision, 20 epochs, seeds 301/302/303. For the three trained arms, ΔH relative to the null is D1 -2.7051, D2 -2.7170, D3 -2.6967 (cross-arm std 0.02). The phenomenon is consistent across probes that differ in parameter count by about two orders of magnitude and differ substantially in architecture: it is a property of the measurement-instrument family, not the RouteScope architecture. Ledger: `experiments/0Z\_refactor\_round/AE30\_minimal\_probe/`.

D.3 Replication without RouteScope (AE42). Fixing $F_1 =$ R1_seed301 and $F_2 =$ R1_seed302 (strict cohort, host-level CKA ≥ 0.989), we train isomorphic probes on the neutral host C1C4_seed123 under C_{task} (segmentation CE, free residual bypass) and C_{recon} (reconstruction MSE, residual bypass zeroed) respectively ($K = 16$, estimator B, the same convention as Table 8), and measure zero-shot $\overline{|H(F_2) - H(F_1)|}$ absolute value per image, then average (three random

Table 7: Replication without RouteScope (AE42).

Probe	Parameter count	$\widehat{D}_F(C_{\text{task}})$	$\widehat{D}_F(C_{\text{recon}})$	Ratio $C_{\text{task}}/C_{\text{recon}}$	95% CI
Linear	6.56×10^4	6.08×10^{-3}	5.72×10^{-4}	$10.65\times$	$[9.65, 11.71]$
MLP	6.58×10^4	5.89×10^{-3}	8.26×10^{-5}	$71.4\times$	$[64.8, 78.4]$

seeds pooled, $n = 768$); the ratio CI comes from an image-level bootstrap ($B = 2000$, seed 42, resampling unit: complete images).

The two families give ratios of different magnitude (the MLP's C_{recon} reading lands on a flatter reconstruction plateau, so its denominator is smaller), but neither CI contains 1, and the criterion is the same: $D_F(C)$ is not invariant to C. A linear probe with no learnable routing already reproduces the core interaction, showing that the interaction is not a product of the RouteScope architecture.

Two convention points require comment. First, the C_{recon} of AE42 is not the same instrument as E2 of Table 8: the reconstruction probe of AE42 has its residual bypass forced to zero ($s = 0$), which presses the reading close to the saturation floor ($H \approx 2.754$, 99.3% of the ceiling), so its $\widehat{D}_F(C_{\text{recon}})$ is closer to the saturation floor of the untrained E3 of Table 8 (2.08×10^{-4}) than to E2 (4.02×10^{-3}). AE42 is therefore a comparison of "one effective probe (C_{task}) against one near-saturated probe (C_{recon})", not the "two effective instruments" comparison of Table 8. Its criterion is whether the CI contains 1 (neither family does), not the magnitude of the ratio. Second, precisely because the C_{recon} probe is already near saturation, the linear family's $10.65\times$ is on the small side (its C_{recon} still has $\widehat{D}_F(C_{\text{recon}}) = 5.72 \times 10^{-4}$, not pressed down to the E3 floor), while the MLP's $71\times$ is on the large side (its C_{recon} has been pressed to 8.26×10^{-5}, below the E3 floor); the magnitude of the ratio varies with the degree of saturation, but saturation affects only the size of the ratio, not its direction or the CI-excludes-1 criterion. The operating-point factor of that panel is 140.6 (identity transform), i.e. extrapolating by operating point gives an order of magnitude more than the measured ratio: the reading differences of that panel fall entirely in the compressive region, which is exactly why it can support only a qualitative criterion. Ledger: `experiments/0Z\_refactor\_round/AE42\_linear\_mlp\_rq2/`.

D.4 Expected differences on the language side (expansion of L1 in Section 7.3). The formalization is modality-agnostic, but three things change on the language side: the probe is usually trained on the very corpus being analyzed, so E overlaps with the measured object and a clean separation of I_0 from E becomes harder; an extra pooling choice (CLS/mean/last token) appears, a fourth factor absent on the vision side; and if the reading switches to an accuracy-like metric, the scale ceiling is set by the task rather than by the instrument, which weakens "operating-point normalization".

E Cross-Backbone, Spatial Analysis and Hyperparameters

E.1 Cross-backbone supporting evidence. ConvNeXt-T (a 2022 pure-convolution model), under the CKA criterion ≥ 0.989 and the same instrument family, gives swap-instrument/swap-representation $= 24.7\times$, CI $[21.9, 28.2]$. The lower bound of the cluster-bootstrap CI of the H ratio being $\geq 10\times$ is the pre-defined criterion, and ConvNeXt-T satisfies it ($21.9\times$). The 4 hosts of that arm are measured under the same E2→E3 range convention as the $117.96\times$ of the R50 strict cohort (F.6); because ratio magnitudes are not comparable across backbones, this is a convention match rather than a second estimate of the same quantity. There is no native E1 on this arm, so what is replicated is the "swap-instrument/swap-representation" range convention, not the $5.93\times$ C_{task}–C_{recon} contrast of Section 4.2. Swin-T did not enter the equivalence cohort because of the cohort structure peculiar to window attention; this is a statement of scope (all claims of this section cover only backbones satisfying that rule), not a negative result for that backbone. Ledger: `experiments/0Z\_refactor\_round/AE26\_crossbackbone\_cohort/`.

E.2 Null-hypothesis calibration of spatial structure. Strict BN panel: $\rho_{\text{entropy}} = 0.7072$, CI $[0.6958, 0.7176]$ (three pairs pooled over 256 images, $n = 768$); the cognate scalars are $R_{l2} = 0.6447$, CI $[0.6317, 0.6570]$, and $R_{\text{maxprob}} = 0.6218$, CI $[0.6082, 0.6343]$. **Original panel**: $\rho = 0.6593$ $[0.6476, 0.6707]$; for both null classes (phase randomization mean ≈ -0.06, circular shift mean ≈ 0) the observed value exceeds all 200 surrogate samples, $p < 5 \times 10^{-3}$. **Per-pixel control**

Table 8: **Measurement condition changes the measured magnitude of the same representational contrasts.** C_{task} and C_{recon} share the same instrument specification but are trained under different objectives. Their host-level paired effect-magnitude ratio is $5.93\times$ (95% CI $[5.71, 6.15]$). C_0 is included only as a saturated boundary condition and is not part of the primary comparison.

Measurement condition	Training environment	Reading baseline H (fraction of ceiling $\ln 16$)	$\widehat{D}_F(C)$ (95% CI)
C_{task} (E1)	task supervision (CE)	1.7146 (61.8%)	2.38×10^{-2} $[1.78, 2.83] \times 10^{-2}$
C_{recon} (E2)	reconstruction MSE pretraining	2.5198 (90.9%)	4.02×10^{-3} $[2.99, 4.84] \times 10^{-3}$
C_0 (E3, boundary)	untrained	2.7608 (99.6%, saturated)	2.08×10^{-4} $[1.51, 2.50] \times 10^{-4}$

panel (no SGLNDP, 3 seeds): $\rho_{\text{entropy}} = 0.9912$, CI $[0.9906, 0.9917]$; the structured-instrument control gives 0.6757, CI $[0.6673, 0.6843]$. Ledgers: `experiments/0Z\_refactor\_round/AE29\_phi\_null/`, `AE34\_pointwise\_control/`.

E.3 Perturbation and behavioral boundary. Across the OOD, Gaussian, JPEG, and salt-and-pepper perturbations, the within-dose per-image correlation has $\max |r| = 0.354$ (the largest of the 18 combinations, $R^2 \leq 0.125$). Conclusion: the response carries structural signal, but is not a direct semantic reading for individual predictions, corresponding to limitation L3 (Section 7.3). Ledger: `experiments/0Z\_refactor\_round/AE17\_nongaussian\_perturbation/`.

E.4 Hyperparameter scan and knob ablation. With $K \in \{1, 4, 8, 16, 32, 64\}$, $l \in \{1, 2, 3, 4, 6\}$, and a scan over C_{state}, $K = 16, l = 3$ is the standard configuration (near the saturation knee). The knob ablation (state bias, self-loop, factor sharpness $\alpha/\beta/\gamma$) is recorded in full; the main text does not treat these knobs as independent contributions.

F Complete Numbers and Convention Distinctions

Table 8: the main comparison of Section 4.2. Every number in the table is stated in Section 4.2; it is placed here to conserve space in the main text.

F.1 Item-by-item readings of Section 4.1. Path-B panel (retrained with `\_eval`, BN frozen): the three models M1/M2/M3 share one frozen backbone and one training environment, with only the instrument independently initialized and trained, and the representations entering the instrument stage are bitwise identical. Readings H: M1 $= 1.1917\,[1.1718, 1.2113]$, M2 $= 0.9975\,[0.9713, 1.0247]$, M3 $= 0.7109\,[0.6894, 0.7330]$, range 0.48, with the three 95% CIs pairwise disjoint. For instruments that differ only in initialization and are untrained, the cross-instrument range of H is 0.00315 (the numerical noise floor). Community-standard panel (original checkpoints, BN updated during training): VOC mIoU 74.21/74.24/73.99 (indistinguishable by prevailing standards); within one measurement condition, the range of H over 5 hosts is 0.0078. **Cross-panel comparison (indicative only, not to be mixed with the numbers above)**: the original-checkpoint E1 (the same recipe as path B's M1, different weights) has a mean reading of 1.7136 over the 5-host strict cohort (AE23), differing from path B's 1.1917 by 0.522, a difference that contains both a weight difference and a host-panel difference and does not constitute a single-factor contrast; panel definitions are in Table 4.

F.2 Table 9: convention distinction between the two "CE vs MSE" contrasts. The same name under different conventions; opposite directions are not a contradiction.

The root cause of the opposite directions lies in the operating point: the first row zeroes the residual bypass on the MSE side, so the reconstruction must flow through the routing distribution P, pushing H toward one-hot (≈ 0.49); the second row's E2 keeps its residual bypass at 1.0, and the reading stays near the ceiling (2.5198). The contribution of the first row is to show that the influence of the training environment (together with the objective-dependent constraints) on the reading is identifiable ($\Delta_E = 1.6537$, Appendix D); the contribution of the second row is to show that two realized instruments give different magnitudes for the same representational difference ($5.93\times$); the

Table 9: Convention distinction between the two "CE vs MSE" contrasts.

Contrast	Instrument instance	Residual bypass s	Training mode	H_{CE}	H_{MSE}	Direction
Operating-point-controlled contrast (Appendix D.1, AE32)	fixed I_0, retrained per objective, 1 host	MSE side set to 0	retrained from scratch per objective	≈ 2.14	≈ 0.49	CE > MSE
Natural operating-point contrast (Section 4.2, AE33)	E1 task-trained / E2 reconstruction-pretrained	≈ 1.35 / 1.0 (natural operating point)	pretrained checkpoint, zero-shot onto 10 hosts	1.7146	2.5198	CE < MSE

two contrasts differ in host panel, residual bypass, and training mode alike, and are not compared under one convention.

F.3 Boundary statement on the cross-panel ordering. The CE/MSE ordering of Appendix D.1 and the E1/E2 ordering of Section 4.2 are opposite on the two panels. The two use different host panels, objective-dependent constraints, and training protocols (the former a controlled environment intervention on community-pretrained hosts, the latter a natural-operating-point contrast on the 10-host R50 panel), so their two trained conditions cannot be compared directly; the untrained level agrees on the two panels (2.7632 vs. 2.7608). This paper therefore makes no claim about the cross-panel ordering of the two trained conditions: "only swapping E reverses the order" should not be read as holding across panels.

F.4 Host panels: construction, strict cohort, and CKA threshold sensitivity. The full panel is a 5×2 design over finetuning recipes and seeds (five learning-rate/weight-decay settings, two seeds each): R1 (lr 3×10^{-3}, wd 10^{-4}, seeds 301/302), R2 (5×10^{-3}, 10^{-4}, seeds 303/304), R3 (3×10^{-3}, 10^{-3}, seeds 305/306), R4 (1×10^{-3}, 10^{-4}, seeds 307/308), R5 (10^{-2}, 10^{-4}, seeds 309/310), giving 10 hosts and $\binom{10}{2} = 45$ pairs. The panel is therefore *not* selected by a similarity threshold, and its pairs are not all mutually indistinguishable: at $\tau = 0.989$ only 13 of the 45 pairs qualify, and the lowest pair is at 0.9545. The strict cohort is the largest mutually indistinguishable subset under the common ruler CKA ≥ 0.989, admitting 5 hosts (R1_seed301/R1_seed302/R2_seed303/R3_seed305/R3_seed306; VOC mIoU tolerance 0.5) and 10 pairs. Re-selecting the cohort at each τ (the same 10×10 CKA matrix, brute-force maximum clique over all 2^{10} subsets): $\tau = 0.989 \rightarrow 5$ hosts (the strict cohort); $\tau = 0.99 \rightarrow 4$ hosts; $\tau = 0.985/0.98 \rightarrow 6$ hosts. The per-host range of E2 at $\tau = 0.989/0.99$ is 2.04×10^{-3} and 1.05×10^{-3}, respectively, always far below $\Delta_C \approx 0.24$, so "representational differences within the indistinguishable interval are amplified hundreds of times by the measurement condition" is robust over $\tau \in [0.98, 0.99]$.

F.5 Conventional similarity readings (AE_COMP01 ledger, backbone GAP features of M1/M2/M3). CKA (pooled over 256 images) 0.9989–0.9992, PWCCA = 1.00000, RSA 0.9985–0.9987, all judged indistinguishable; the same ledger's instrument reading is $H = 1.6981/0.6003/1.1987$ (range 1.0978). Ledger: `experiments/0Z\_refactor\_round/AE\_COMP01\_traditional\_tools/ae\_comp01\_result.json`.

Panel attribution. AE_COMP01 loads the original checkpoints `C1C4\_seed\{123,124,125\}/best\_model.pt` (community-standard panel, BN statistics updated during training and differing between models), with similarity measured on backbone GAP features; the H triple of its `routescope\_reference`, $1.6981/0.6003/1.1987$, is bitwise identical to `H\_C4` of `experiments/0E\_A\_checkpoint\_fingerprint/matched\_performance\_summary.json`, that is, the reading of the original-checkpoint instrument on community-pretrained hosts. The $1.1917/0.9975/0.7109$ of the path-B panel in Section 4.1 come from _eval retraining of the same-named recipe (`--eval-backbone`, BN frozen): readings from the two panels cannot be mixed, and the difference comes from both the weights and the BN protocol, which is expected rather than a defect of provenance. The single value 1.7136 of F.1 is the mean of the original-checkpoint E1 over the 5-host strict cohort (AE23); each of the three readings belongs to its own panel, as noted throughout this section.

F.6 The flagship-aligned 2×2 four-cell table (the single pair R1_seed301 vs. R1_seed302, per-image paired cluster bootstrap, $n = 256$): representation main effect $\Delta_F = -5.87 \times 10^{-4}$, CI $[-9.18, -2.54] \times 10^{-4}$; interaction $\Delta_{FC} = -6.02 \times 10^{-4}$, CI $[-9.30, -2.75] \times 10^{-4}$ ($\Delta_F(\mathrm{E2})$ and $\Delta_F(\mathrm{E3})$ have opposite signs, hence subtraction rather than addition). The measurement-condition main effect is $\Delta_C = 0.2411$, CI $[0.2389, 0.2433]$. Raw four-cell table: $R(F_1, \mathrm{E2}) = 2.519619$, $R(F_2, \mathrm{E2}) = 2.520206$, $R(F_1, \mathrm{E3}) = 2.760823$, $R(F_2, \mathrm{E3}) = 2.760808$, yielding $\Delta_F(\mathrm{E2}) = -5.87 \times 10^{-4}$, $\Delta_F(\mathrm{E3}) = +1.5 \times 10^{-5}$ (opposite signs), $\Delta_{FC} = -6.02 \times 10^{-4}$, $\Delta_C = 0.241204$. The range convention gives swap-instrument/swap-representation $= 117.96\times$, CI $[95.79, 152.50]\times$.

G REPRODUCIBILITY

Ledgers and scripts. Every load-bearing number corresponds to a named ledger (`experiments/0Z\_refactor\_round/AE*/`) and a generating script (`scripts/refactor\_ae*.py`); numbers, fields, bootstrap parameters, and convention names are bound item by item and guarded with `assert`; historical readings that cannot be recomputed automatically are flagged in the ledger as `MANUAL`/`CITED` together with an alternative recomputation path. The main-panel pairwise effect magnitudes, the ordering and the reversal count ($7/45$, 0.9833) are directly re-computable from the three JSON files of `AE33\_hero\_fullpanel/` (reproduced bitwise in the guards of `refactor\_ae44\_hero\_scatter.py`). Table 1 lists seeds, checkpoint identifiers, and training recipes; Appendices A and D present the configurations for the three intervention classes.