SAME SCORE, SAME CAPABILITIES? REGULARIZING FACTORIZED REPRESENTATIONS FOR TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Similar scores need not imply equally reusable components. We ask whether the selection readout should match the representation interface used downstream. A population construction shows that identical full-representation validation, transfer and spectra can conceal different component utility. FACTOR-MAE controls separate weight-decay strength from same-strength dependence regularization. **Utility-Constrained Regularization Selection (UCRS)** screens strength, then maximizes backbone transfer subject to unique-code non-decrease. Separate selection and evaluation cohorts test whether this constraint improves downstream component reuse. Exchanging code and trunk guards reverses preference across student targets, with a 0.250-point interaction. A second, prediction-preserving interface intervention yields reciprocal preferences in both ViT-Ti and ResNet18, with interactions $0.313/0.312$ points. These results support matching the selection readout to the consumer's interface, rather than simply adding a probe. Unique-only tradeoffs and retained student losses delimit the benefit: UCRS is neither student-optimal nor a guarantee of loss-free reuse. The anonymized code, data registry and experiment records are available at `https://anonymous.4open.science/r/ID20229`.

1 INTRODUCTION

When two models achieve similar validation scores, have they learned equally useful representations? Just as equal exam scores need not imply identical abilities, agreement on one evaluation need not imply that two models retain the same useful information. This distinction matters for unified multimodal models (UMMs), which combine understanding and generation (Xie et al., 2025; Chen et al., 2025; Li et al., 2026), and it compounds as such systems are assembled and reused: a representation validated under one readout is routinely frozen as the backbone, or distilled as the teacher, of a second model. Their representations must support reusable semantic information while retaining details needed by individual tasks. A validation score measures success under one readout; it does not establish whether the representation remains useful when that readout, task, or training environment changes. A regularization choice made through that same score inherits the blind spot. How a representation behaves under one readout therefore need not predict how it behaves under the interface a downstream consumer will actually attach.

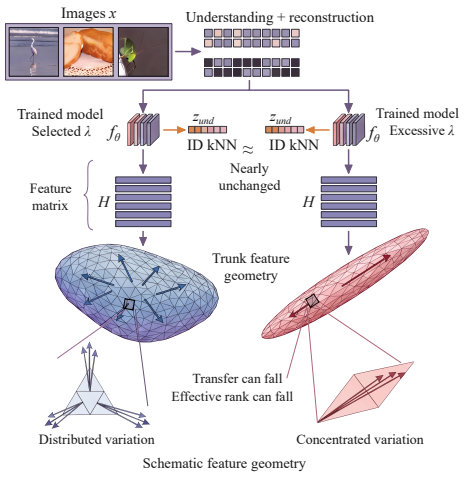

Figure 1: **Similar scores, different utility.** Near-matched in-domain accuracy can conceal different transfer and feature geometry. Thumbnails and geometry are illustrative.

Factorization makes the access to that information explicit. Shared and unique codes are intended to separate reusable and task-specific information, using mutual-information (MI) objectives or orthogonality constraints alongside ordinary weight decay (Lu et al., 2026; Estepa et al., 2026). A downstream head may access the backbone, whereas a lightweight model may learn from a designated

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.

code target. Good performance through the first interface does not certify the second. Our question is therefore not simply whether a second score helps: *which component's readout should constrain regularization for a specified consumer?* Code names and low dependence alone do not establish that correspondence.

We study this question in FACTOR-MAE, whose understanding and masked-reconstruction streams share a ViT backbone. This is a visual testbed for UMM-inspired factorization, not a complete multimodal generator. We separate two interventions: choosing weight-decay strength, then adding MI at that same strength. We distinguish in-domain accuracy, frozen-backbone transfer and unique-code discriminative utility; spectra describe geometry without certifying usefulness. Capabilities refer only to these readouts. This design isolates the reported ON/OFF contrast from differences in the weight-decay reference, without claiming that a probe certifies all uses of a component.

The measured weight-decay sweep illustrates this discrepancy (Fig. 1). From the selected to an excessive strength, ViT-B in-domain accuracy changes from 19.05 to 18.79, while transfer falls from 49.49 to 47.03 and effective rank drops from 537.3 to 443.6. More subtly, a penalty that helps backbone transfer can affect unique-code utility differently. Neither outcome subsumes the other. A reusable backbone and a specialized code have distinct responsibilities: accepting one improvement need not justify sacrificing the other. A population construction sharpens this distinction. Merely exchanging coordinates preserves full-representation validation, transfer and covariance, yet changes the utility of a fixed exposed component. The issue is not just how much information survives, but whether it remains accessible through the interface that will be used. This existence result motivates evaluating the interface; it does not explain learned factorization.

Utility-Constrained Regularization Selection (UCRS) instantiates this question for backbone reuse with a unique-code utility requirement. Short-budget shared-code kNN ranks weight-decay strengths. At the selected strength, matched full-budget models with and without MI estimate the conditional ON/OFF contrast. UCRS maximizes transfer subject to measured code non-decrease against calibrated WD; it does not claim to protect every default capability. Screening is shared by the competing rules. The scientific question concerns the constraint's information: does the code readout distinguish interventions that matter to a code-target consumer, beyond what backbone transfer or another probe reveals? Replacing the code guard with a matched trunk guard tests this hypothesis without changing candidates, reference strength or the transfer objective. ID accuracy and spectra remain diagnostic.

The weight-decay study establishes the validation–utility gap and selected-strength controls, but its six retrospective choices coincide with transfer-only and its primary-suite gain reverses on the broader suite. We therefore test the incremental claim with separate selection and evaluation cohorts. Across 9 conditions, UCRS changes 3 transfer-only decisions. Exchanging code and trunk guards reverses preference across student targets: code guarding gains 0.124 points for code targets, whereas trunk guarding gains 0.126 for trunk targets. A second control changes the exposed code interface while preserving each candidate's original task predictions. Reciprocal preferences persist in both ViT-Ti and ResNet18; complete-selection statistics assess both directions, not merely a positive interaction. These controls test target usefulness, not universal factorization benefit or guaranteed distillation quality. False rejection, retained losses and default-relative student degradation remain.

Our contributions are threefold:

- **A distinction between whole-representation and component utility.** Measured controls isolate same-strength interventions; a population construction shows why even full-readout agreement cannot certify a designated component.
- **A matched-reference method for component reuse.** UCRS optimizes backbone transfer under an explicit code-utility constraint. Its contribution is the reuse-specific choice of evaluation interface, not a new generic constrained optimizer.
- **Evidence for consumer-matched selection.** Crossed guard/target comparisons, a prediction-preserving interface intervention and two students distinguish matching from generic caution. Complete-selection statistics quantify these contrasts; failures and direct competing rules delimit the claim.

2 RELATED WORK

Unified understanding/generation models use diverse visual encodings (Chameleon Team, 2024; Zhou et al., 2025; Xie et al., 2025; Chen et al., 2025). DIVA studies shared/unique factorization (Lu et al., 2026), LEASE uses discriminative codebook supervision (Estepa et al., 2026), and SPAR couples semantic–pixel alignment with adaptive routing (Li et al., 2026). DualToken separates semantic and reconstructive visual codebooks (Song et al., 2026); RecA conditions reconstruction on visual-understanding embeddings (Xie et al., 2026). These motivate our image-only testbed, not a reproduced UMM. SSL objectives and variance/covariance constraints (Chen et al., 2020; Grill et al., 2020; Zbontar et al., 2021; Bardes et al., 2022; Jing et al., 2022) address degeneration; MI bounds and CLUB (van den Oord et al., 2018; Belghazi et al., 2018; Poole et al., 2019; Cheng et al., 2020) regularize dependence. Neither low dependence nor high effective rank (Roy & Vetterli, 2007) certifies frozen-feature transfer (Kornblith et al., 2019; Ericsson et al., 2021).

Low-budget selection is established by successive halving and Hyperband (Jamieson & Talwalkar, 2016; Li et al., 2017); constrained and multiobjective optimization also predate UCRS (Gardner et al., 2014; Hernández-Lobato et al., 2016). In particular, Learn then Test and Pareto Testing (Angelopoulos et al., 2021; Laufer-Goldshtein et al., 2023) select utility under calibrated risks, including loss relative to a reference. UCRS is not a new generic optimizer or risk test. Its proposed contribution is the component-reuse instance: shared-code screening followed by a same-strength intervention evaluated through distinct backbone/code interfaces. The claimed value must come from useful additional decision information, not the mere presence of two stages or a constraint.

3 UTILITY-CONSTRAINED REGULARIZATION SELECTION

Testbed. FACTOR-MAE adapts shared/unique factorization (Lu et al., 2026) and codebook supervision (Estepa et al., 2026). A randomly initialized ViT (Dosovitskiy et al., 2021) processes heavily masked reconstruction (He et al., 2022) and lightly masked understanding streams. Gated MLPs yield

$$z_g^{\mathrm{sh}} = E_{\mathrm{sha}}(f_g), \quad z_u^{\mathrm{sh}} = E_{\mathrm{sha}}(f_u), \quad z^{\mathrm{gen}} = E_{\mathrm{gen}}(f_g), \quad z^{\mathrm{und}} = E_{\mathrm{und}}(f_u). \tag{1}$$

Heads read $[z_g^{\mathrm{sh}}; z^{\mathrm{gen}}]$ for masked pixels and $[z_u^{\mathrm{sh}}; z^{\mathrm{und}}]$ for DINOv2-derived codebook targets (Oquab et al., 2024); initialization is random, but supervision is not teacher-free. The objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \mathcal{L}_{\mathrm{disc}} + w_{\mathrm{sh}}\mathcal{L}_{\mathrm{InfoNCE}}(z_g^{\mathrm{sh}}, z_u^{\mathrm{sh}}) + w_{\perp}\mathcal{L}_{\perp} + w_{\mathrm{mi}} \sum_{a\in\{g,u\}} \hat{I}(z_a^{\mathrm{sh}}, z_a^{\mathrm{unique}}). \tag{2}$$

Here $z_g^{\mathrm{unique}} = z^{\mathrm{gen}}$, $z_u^{\mathrm{unique}} = z^{\mathrm{und}}$, $\mathcal{L}_{\perp}$ is shared–unique squared cosine and vanilla CLUB penalizes within-flow dependence. This differs from DIVA's cross-flow unique–unique NCE-CLUB objective. Factor-off omits unique encoders and mechanism terms, with task heads reading shared codes only. CLUB-off adds unique encoders, shared InfoNCE and orthogonality; full additionally enables the MI term.

Distinct readouts. ID accuracy uses unique-understanding-code kNN; screening instead uses shared-code kNN. Transfer (T) uses frozen-trunk linear probes with at most 100 labels/class; unique-code utility (U) uses the same probes on z^{und}. The primary suite equally weights CIFAR-100, DTD and Food-101; the broader suite adds Pets, Flowers and Cars. Geometry uses centered singular values: $p_i = \sigma_i / \sum_j \sigma_j$, $r_{\mathrm{eff}} = \exp(-\sum_i p_i \log p_i)$ and $\kappa = p_1$. These diagnostics are not selection thresholds. Similar ID scores are numerically close observations, not statistical equivalence; changing readout and task does not isolate domain shift.

3.1 WHAT VALIDATION AGREEMENT LEAVES UNDETERMINED

Proposition 1 (Equal scores and spectra, different transfer utility)**.** *Let X, U be independent uniform signs and let $N \sim \mathcal{N}(0,1)$ be independent of both. For $t \geq 0$, define*

$$Z_t = \left(X, \frac{tU + N}{\sqrt{1+t^2}} \right). \tag{3}$$

For validation target X and transfer target U, optimal population accuracies are $\mathcal{A}_{\mathrm{id}}^* = 1$ and $\mathcal{A}_{\mathrm{tr}}^* = \Phi(t)$ (standard normal CDF), both attained by linear classifiers. Yet $\mathrm{Cov}(Z_t) = I_2$: singular-value effective rank stays 2 and leading-component share stays $1/2$.

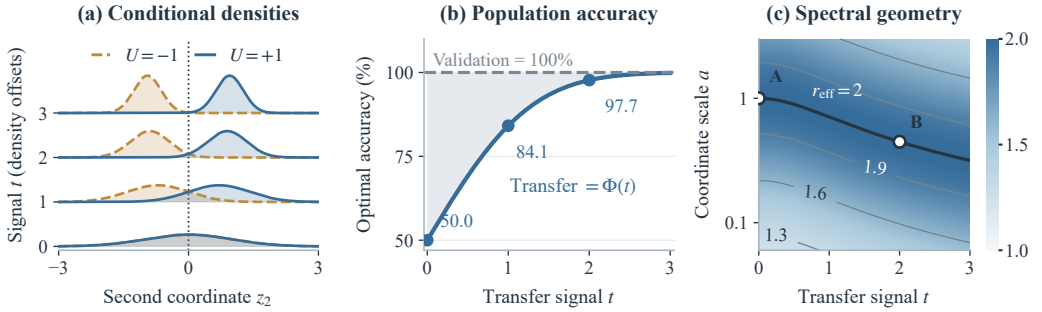

Figure 2: **An analytic view of validation equivalence.** **(a)** Exact z_2 class-conditional densities, vertically offset for four values of t. **(b)** Transfer improves at fixed validation accuracy. **(c)** Effective rank of $(X, a(tU+N))$; the dark curve $a = (1+t^2)^{-1/2}$ keeps covariance I_2, $r_{\text{eff}} = 2$ and $\kappa = 1/2$. A and B share scores and spectra, but attain 50.0% and 97.7% transfer accuracy. Population quantities, not trained-model measurements.

Figure 2 varies discriminability without changing variance. There is also an *interface blind spot*: swap the coordinates of Z_t. Full-representation validation, transfer and spectra are unchanged, yet predicting U from the fixed second coordinate drops from $\Phi(t)$ to $1/2$. Thus even agreement on whole-representation transfer cannot certify component utility. This population counterexample is not a claim about learned factorization or student behavior: the swap preserves the full readout, while independence of X and U fixes the exposed component at chance.

3.2 The UCRS rule

UCRS chooses weight decay λ and whether to add the specified MI weight; architecture and task objectives remain fixed (Fig. 3).

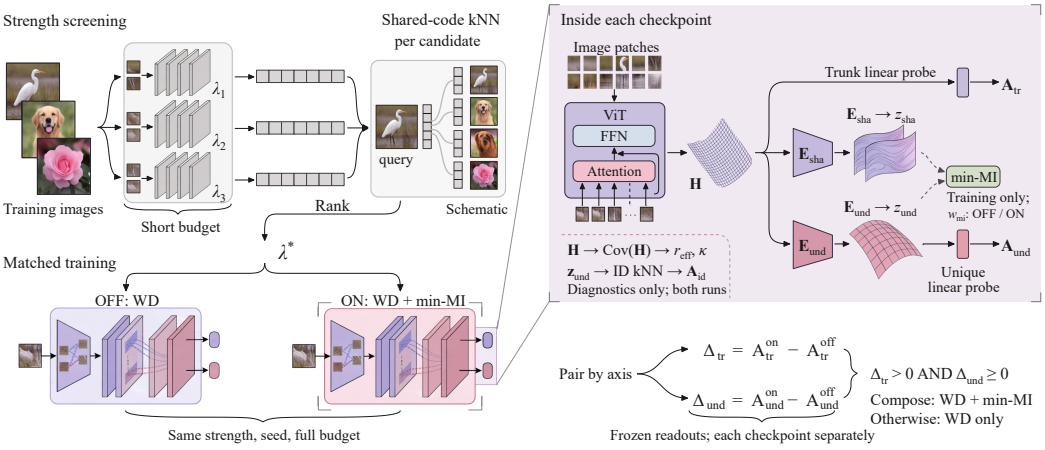

Figure 3: **UCRS: screening and utility-constrained selection.** Shared-code kNN screens strengths. At full budget, transfer from H is the objective; unique-code utility from z_{und} supplies the non-decrease constraint (Eq. 6). The shown AND rule is its two-candidate solution. ID and spectra are diagnostic. Task heads and reconstruction are omitted; photographs, neighbors and geometry are illustrative.

Screen strength. Train each MI-OFF candidate for bE epochs, where E is the full horizon. Shared-code validation kNN selects

$$\lambda^\star \in \arg\max_{\lambda \in \Lambda} s_b(\lambda). \tag{4}$$

Screening ties choose the smallest strength. This ranks candidates, not calibrated transfer quality. At $\lambda^\star$, train full-budget OFF/ON arms with matched data, architecture, horizon, non-MI losses, seed pairs and probe protocol. The nonzero MI weight is prescribed, not tuned on evaluation.

Compose subject to component utility. Let $\mathcal{A}_{\mathrm{tr}}, \mathcal{A}_{\mathrm{und}}$ be empirical backbone/code probe accuracies and define

$$\Delta_{\mathrm{tr}} = \mathcal{A}_{\mathrm{tr}}^{\mathrm{ON}} - \mathcal{A}_{\mathrm{tr}}^{\mathrm{OFF}}, \qquad \Delta_{\mathrm{und}} = \mathcal{A}_{\mathrm{und}}^{\mathrm{ON}} - \mathcal{A}_{\mathrm{und}}^{\mathrm{OFF}}. \tag{5}$$

UCRS solves the two-candidate problem

$$w_{\mathrm{mi}}^{\star} \in \arg \max_{w \in \{0, w_{\mathrm{mi}}\}} \mathcal{A}_{\mathrm{tr}}(\lambda^{\star}, w) \quad \text{s.t.} \quad \mathcal{A}_{\mathrm{und}}(\lambda^{\star}, w) \geq \mathcal{A}_{\mathrm{und}}(\lambda^{\star}, 0). \tag{6}$$

To derive the decision, write $w = a w_{\mathrm{mi}}$ with $a \in \{0, 1\}$. Subtracting the same-strength WD reference from the objective and constraint gives

$$a^{\star} \in \arg \max_{a \in \{0,1\}} a\Delta_{\mathrm{tr}} \quad \text{s.t.} \quad a\Delta_{\mathrm{und}} \geq 0, \tag{7}$$

$$a^{\star} = \begin{cases} 1, & \Delta_{\mathrm{tr}} > 0 \text{ and } \Delta_{\mathrm{und}} \geq 0, \\ 0, & \text{otherwise}, \end{cases} \qquad w_{\mathrm{mi}}^{\star} = a^{\star} w_{\mathrm{mi}}. \tag{8}$$

OFF is always feasible; ON is excluded by a utility loss and otherwise wins only with a strict transfer gain. The reduction is exact for two candidates, not an interpolation over MI weights. Missing readouts mean insufficient evidence. For n paired seeds, $\Delta_k = n^{-1} \sum_s (\bar{A}_{k,s}^{\mathrm{ON}} - \bar{A}_{k,s}^{\mathrm{OFF}})$, where $\bar{A}$ averages probe repeats within a seed and $k \in \{\mathrm{tr}, \mathrm{und}\}$. Unequal historical cohorts are not paired. The constraint protects measured utility relative to calibrated, same-strength WD, not default-model utility or all component information.

With m screening runs and r full-budget seed pairs, nominal training cost is $mb + 2r$ full-run epoch-equivalents, excluding probes and evaluation. The $1/5$ fraction applies per screening candidate, not end to end. Neither accounting nor feasibility supplies significance or a shift guarantee.

4 Evidence and Analysis

Measured protocol. Randomly initialized ViT-Ti/S/B/L models train on IN-100 for 200 epochs; ViT-S/B also use IN-1k for 30 epochs. Trunks remain trainable during pretraining and are frozen only for downstream evaluation. The primary and broader suites use the same fixed task weights within a comparison. Tables identify available-seed means, SDs and readout-specific cohorts; unequal arms are not paired. Selection inputs and diagnostic measurements are distinguished from the independent evaluation outcomes in Section 4.3.

4.1 Validation agreement conceals utility changes

Across six ViT-B strengths, ID kNN spans only 0.26 pp (Fig. 4). Increasing WD from 0.4 to 0.8 lowers transfer from 49.49 to 47.03, unique-code accuracy from 35.32 to 30.35, and effective rank from 537.3 to 443.6. The unique-code comparison uses two seeds versus one, so its 4.97-point loss is descriptive, not paired. Concentration rises from 0.0159 to 0.0211; this measures spectral shape, not a percentage of information lost.

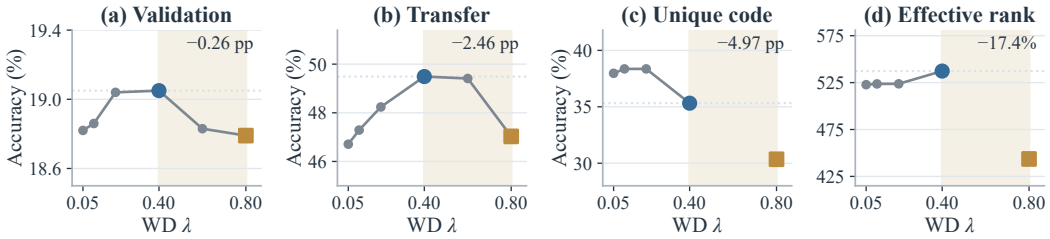

Figure 4: **Similar scores, different capabilities.** ViT-B/IN-100, 200 epochs; four readouts share the strength axis. Blue/gold mark WD 0.4/0.8. The unique-code mean at 0.4 uses two seeds, versus one at 0.8; shading marks the comparison, not uncertainty. Missing observations at 0.6 remain gaps.

Table 1: **Same-strength comparisons across six measured settings.** Available-seed means and SDs (%); Δ is same-strength MI-minus-WD (pp). Shaded rows are retrospective UCRS choices.

		Transfer: 3-suite			Transfer: 6-suite			$\mathcal{A}_{\text{id}}$			$\mathcal{A}_{\text{und}}$			
Configuration	λ	Mean	SD	Δ	Mean	SD	Δ	Mean	SD	Δ	Mean	SD	Δ	n
ViT-B / IN-100							UCRS: retain MI							
WD at λ^*	.4	49.51	0.16	—	52.27	0.28	—	19.08	0.07	—	35.36	0.55	—	4
UCRS: WD+MI	.4	50.11[¶]	0.15	+0.60[¶]	52.01	0.42	−0.27	19.10	0.03	+0.03	36.30	0.42	+0.94	4
ViT-Ti / IN-100							UCRS: keep WD							
UCRS: WD	.4	40.41	0.09	—	41.79	0.12	—	17.09	0.16	—	36.03[§]	—	—	2
WD+MI at λ^*	.4	39.21	0.20	−1.20	40.07	0.04	−1.72	16.72	0.02	−0.38	35.35[§]	—	−0.68	2
ViT-S / IN-100							UCRS: retain MI (small, uncertain)							
WD at λ^*	.4	45.70	0.18	—	48.09	0.14	—	18.36	0.23	—	34.47	0.65	—	3
UCRS: WD+MI	.4	45.79	0.15	+0.09	47.72	0.22	−0.37	18.53	0.06	+0.17	36.14*	0.60	+1.67	3
ViT-L / IN-100							UCRS: keep WD							
UCRS: WD	.4	51.15[†]	0.01	—	53.64[†]	0.07	—	19.33	0.06	—	38.51	0.24	—	3
WD+MI at λ^*	.4	50.40	0.31	−0.74	51.53	0.18	−2.11	18.95	0.09	−0.38	39.05	0.55	+0.54	3
ViT-S / IN-1k, 30 epochs							UCRS: keep WD (transfer declines); kNN gain							
UCRS: WD	.1	65.51	0.34	—	65.25	0.58	—	47.23	0.15	—	60.80	0.55	—	3
WD+MI at λ^*	.1	65.42	0.41	−0.09	64.79	0.28	−0.46	47.72	0.21	+0.48	61.40	0.67	+0.60	3
ViT-B / IN-1k, 30 epochs							UCRS: retain MI (n=2)							
WD at λ^*	.1	72.15	0.01	—	71.89	0.16	—	54.05	0.52	—	65.52	0.06	—	2
UCRS: WD+MI	.1	72.31	0.06	+0.17	71.61	0.14	−0.28	54.24	0.27	+0.20	67.46	0.12	+1.94	2

n: training seeds; exceptions: *unique $n = 2$, [†]transfer $n = 2$, [§]unique $n = 1$ (no SD). S unique OFF/ON (3/2) and L transfer (2/3) are unpaired. [¶]B ON seed 1 differs across registry/six-task sources (primary margins $+0.60/+0.46$ pp). Early-ranking validation is B/IN-100 only. Deltas precede display rounding.

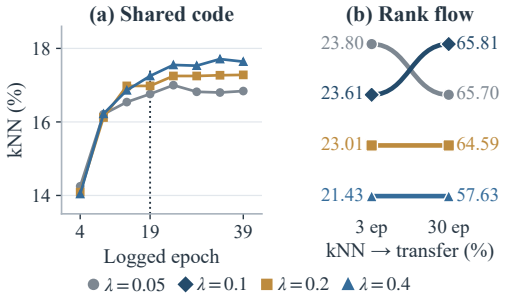

Figure 5: **Early screening across two environments. (a)** ViT-B shared-code kNN; epochs are zero-based. **(b)** ViT-S/IN-1k: 3-epoch kNN versus 30-epoch six-task transfer.

Screening is environment-specific. ViT-B's 40-epoch shared-code probe recovers the 200-epoch transfer ordering of $\{0.05, 0.2, 0.4\}$; ranking stabilizes at logged epoch 19. The 3-of-30-epoch ViT-S/IN-1k probe instead swaps the near-tied leaders $0.05/0.1$. Its reported 0.1 choice uses a separate full-budget comparison, not the early argmax. Other settings test composition, not screening recovery. Thus the tested $1/5$ screen budget does not establish an environment-independent selection guarantee. Early-score stability and agreement with the eventual transfer ordering are distinct checks. The latter must be established locally before treating a cheaper screen as a substitute for full-budget strength comparison.

Ordinary regularization is a necessary control. Default-strength CLUB improves ViT-B transfer by +1.15 points, while ordinary WD gains +1.04 and +2.29 at $0.2/0.4$ over its matched-environment default. Thus an MI gain needs a calibrated WD reference. Among critic-free controls, nHSIC reduces dependence most but transfers worst; CLUB has the smallest average cross-correlation reduction yet the highest transfer/code utility. Estimator and optimization effects differ: dependence reduction does not order utility.

4.2 Measured same-strength comparisons and their limits

Table 1 retains the six selected-strength comparisons. UCRS choices are retrospective applications to available arms, not independently validated end-to-end selections.

ViT-B's four-pair primary gain is $+0.60$ pp in the registry versus $+0.46$ in the six-task records (ON seed 1: $50.00/49.42$); each is analyzed separately. Broader-suite means reverse to $-0.265/-0.263$ pp. Food-101 improves in every pair; Pets, Flowers and Cars each decline, averaging -0.983 pp. Leave-one-pair-out analysis preserves these signs (Fig. 7). Four tasks outside both suites yield +0.35 pp for frozen-trunk reuse, with a Caltech-101 decline. This tests a configuration, not the constraint: transfer-only and UCRS choose identical checkpoints in all six settings. Ti/L losses, S uncertainty and IN-1k broader-suite reversals preclude unconditional improvement.

Geometry is complementary, not sufficient. Figure 6 combines layerwise ranks, full spectra and 1,024 paired image readouts. Excessive WD concentrates singular-value mass and increases image-mean patch cosine by median 0.190; default-strength MI has a different depth profile. These are descriptive changes, not causal mechanisms or extra feasibility conditions.

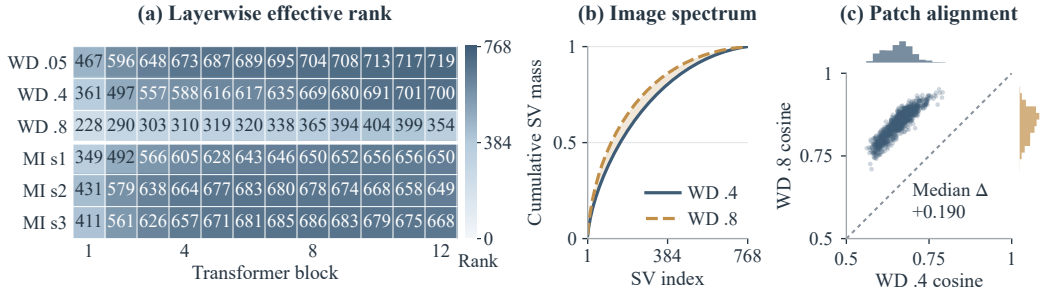

Figure 6: **Feature concentration across layers, images and patches. (a)** Pre-final-normalization token rank, six runs and twelve blocks; cells show rounded ranks. WD rows use seed 1; MI rows use $\lambda = .05$. **(b)** Cumulative normalized singular-value mass from all 768 pooled components; shading shows the difference between the two curves, not uncertainty. **(c)** One point per image compares its mean patch-to-image cosine, with identical marginal histogram bins. Panels (b,c) use one seed-1 checkpoint per strength and a common FP32 extraction; all panels share image identities.

4.3 When does the component constraint change a useful choice?

Shared candidates, separate evaluation. ViT-B/IN-100, ViT-S/IN-100 and ViT-S/IN-1k each use three prescribed MI weights, giving nine conditions. Rules share 24 full-budget candidates, WD references, task weights and probe supports. Selection and evaluation cohorts are separate; each rule is fixed before evaluation. A ViT-Ti learns normalized code targets from the teacher while seeing raw images. Its accuracy S averages CIFAR-10, SUN397, Caltech-101 and EuroSAT and never enters selection.

Unit of repetition. Complete-selection statistics use $n = 24$ outer blocks. Each restarts all three screening grids, selects strengths and fits matched OFF/ON teachers before independent evaluation. Two student initializations per target and three support draws per student are averaged within blocks; neither these repeats nor the nine conditions increase n. Rules reuse candidate outcomes within each block, so comparisons are paired.

Table 2: **Changed choices and their student consequences.** Separate selection and evaluation readouts. Contrasts use TO or the named WD reference; final $T/U/S$ columns report UCRS accuracies.

		Selection			Choice				UCRS minus TO (pp)			Student vs WD		UCRS evaluation (%)		
Case	β	d_T	d_U	d_k	UCRS	TO	UO	V	δ_T	δ_U	δ_S	Local	Default	T	U	S
H301	0.03	+0.25	+0.45	-0.01	M	M	M	M	0.00	0.00	0.00	+0.20	-0.45	82.30	35.75	75.40
H302	0.1	+0.60	+0.94	+0.02	M	M	M	W	0.00	0.00	0.00	+0.40	-0.25	82.45	36.10	75.60
H303	0.3	+0.35	−0.80	+0.05	W	M	W	M	−0.25	+1.10	+0.65	0.00	-0.65	82.10	35.40	75.20
H304	0.03	+0.15	+0.25	-0.01	M	M	M	M	0.00	0.00	0.00	+0.10	-0.25	78.90	31.35	71.50
H305	0.1	+0.30	−0.55	+0.04	W	M	W	M	−0.20	+0.85	+0.45	0.00	-0.35	78.80	31.20	71.40
H306	0.3	−0.10	+0.25	+0.08	W	W	M	W	0.00	0.00	0.00	0.00	-0.35	78.80	31.20	71.40
H307	0.03	+0.20	+0.10	+0.01	M	M	M	W	0.00	0.00	0.00	-0.04	-0.04	80.70	60.62	78.06
H308	0.1	+0.25	−0.08	+0.00	W	M	W	M	−0.15	−0.05	-0.04	0.00	0.00	80.60	60.67	78.10
H309	0.3	−0.05	−0.45	+0.02	W	W	W	M	0.00	0.00	0.00	0.00	0.00	80.60	60.67	78.10
Equal-condition mean									−0.07	+0.21	+0.12	+0.07	−0.26	80.58	42.55	74.97

$\beta = w_{\mathrm{mi}}$; d_k: MI-minus-WD unique kNN (pp); W/M: WD/MI; TO/UO/V: transfer-only/unique-only/validation-only. V may change strength. Gray rows change choices. H301–303/E1, H304–306/E2, H307–309/E3 share grids/WD arms. Local/default: same-strength/default WD; rules: Table 5. Final $T/U/S$: selected-model transfer/unique LP/student accuracy; evaluation-only. Conditions are not seeds or significance tests.

The constraint changes three decisions: H303/H305 preserve $0.65/0.45$ student points at a transfer cost, whereas H308 is a false rejection costing 0.04 student points. H307 retains MI but loses 0.04 points relative to its own WD teacher. Averaged UCRS-minus-transfer-only outcomes are -0.0667 transfer, $+0.2111$ code and $+0.1178$ student points.

4.4 MATCHING THE GUARD TO THE CONSUMER

Exchange the guard and the target. T+A replaces the code guard with $d_A \geq 0$, probing a fixed 256-dimensional trunk projection; $d_T > 0$, screening, candidates, supports and ties stay fixed. Each teacher supplies both code and trunk student targets. This tests whether the useful guard follows the consumer's interface.

Table 3: **Does the useful guard depend on the student target?** Crossed-control outcomes.

Readout	TO	UCRS	T+A	UO
MI choices	7	4	5	5
Code $S_C \uparrow$	74.856	<u>74.973</u>	74.849	**74.996**
Trunk $S_A \uparrow$	<u>74.673</u>	74.603	**74.729**	74.592
D_C	+0.000	+0.118	-0.007	+0.140
D_A	+0.000	-0.070	+0.056	-0.081
$D_C - D_A$	+0.000	+0.188	-0.062	+0.221
WD choices	2	5	4	4
Constraint	None	$d_U \geq 0$	$d_A \geq 0$	$d_U > 0$

Bold/underline: best/runner-up per readout, not significance. UO: unique-only. MI/WD count choices among nine dependent conditions. D_C, D_A: gains over TO. T+A: dimension-matched trunk guard. All but UO require positive transfer. Rules reuse candidates; conditions are not replicates.

UCRS exceeds T+A by .124 points on code targets; T+A exceeds UCRS by .126 on trunk targets. Their sum, .250, is the guard–target interaction, not either student's accuracy gain. Both directional preferences support matching. The paired 95% intervals are $[.066, .183]$ and $[.051, .200]$ under the four-contrast Bonferroni family. T+A and unique-only accept equally many MI arms but favor different targets: acceptance frequency alone does not explain utility. UCRS retains 4 arms versus transfer-only's 7; H307 still loses student utility despite its positive interaction contribution.

Change the interface, not the teacher. A second control exposes u or $q = (u+s)/\sqrt{2}$, where s is the shared code. Inverse orthogonal mixing before the original heads preserves each teacher's predictions and paired-code covariance spectrum. Only the guarded interface changes between G_u (UCRS) and G_q; six choices differ. For ViT-Ti, matching gains .169 points on u and .144 on q; ResNet18 gains $.160/.152$ (Table 4). Both directions remain positive: the $.313/.312$ interactions summarize reciprocal preferences, rather than single-model gains. Thus preference follows the exposed interface even with original task predictions fixed; guards may still select different teachers.

Table 4: **Fixed-teacher interface crossing.** Complete-selection contrasts, in pp.

Target	Mean	SD	Lower	Upper	n
ViT-Ti					
U	+0.169	0.120	+0.098	+0.240	24
Q	+0.144	0.120	+0.074	+0.215	24
$U+Q$	+0.313	0.196	+0.198	+0.429	24
ResNet18					
U	+0.160	0.120	+0.089	+0.231	24
Q	+0.152	0.120	+0.082	+0.223	24
$U+Q$	+0.312	0.196	+0.197	+0.428	24

U: $G_u - G_q$ on the u target; Q: $G_q - G_u$ on $q = (u+s)/\sqrt{2}$. Interaction is their sum. SD and n summarize outer blocks. Bounds use the same six-contrast Bonferroni family; conditions are not replicates.

Precision and intended use. Paired Student-t intervals assume independent, approximately normal outer contrasts. Bonferroni 95% families cover four primary contrasts, six interface contrasts and two unique-only contrasts separately. Unique-only exchanges .022 backbone points for .022 more student points through H306, with paired intervals $[.008, .037]$ on each side. UCRS therefore serves backbone transfer with a code-utility requirement; it does not optimize student accuracy alone.

Limits under a different objective. For $T + wS$, continuous-threshold oracle gaps are only $0.004387/0.003429$ weighted points at $w = .6$ (excluding noisy T+A). Default WD better serves student-only use. Application thresholds and costs are illustrative; these small gaps do not establish deployment value.

Beyond a stricter transfer threshold. Table 5 shows fewer code losses than transfer-only (three to one; maximum 1.10 to .05 pp), at a transfer cost. UCRS accepts H301 at $d_T = 0.25$ but rejects H303 at 0.35; it accepts H304 at 0.15 but rejects H305 at 0.30. No monotone transfer threshold reproduces these inversions. Across all 64 environment-specific acceptance sets, UCRS exceeds their maximum code/student means by $.0444/.0244$ pp. This oracle comparison isolates decision information beyond a transfer threshold; it is not a trained baseline.

Changing the reference changes the question. With fixed task weights, changing from screened to default WD adds the same reference offset to every rule's gain. UCRS's student contrasts are therefore $+0.073$ and -0.260 points, respectively. The constraint protects composition at the screened strength; it cannot undo earlier strength-selection losses. Both references distinguish local improvement from end-to-end preservation.

Table 5: **Transfer gains and component-use losses.** Losses use shared-screen WD; gains show both calibrated and default WD references.

	Selection information			Gain over screened WD (pp)			Code loss		Student loss		Gain over default WD (pp)		
Rule	Screen	MI gate	MI	$\Delta T \uparrow$	$\Delta U \uparrow$	$\Delta S \uparrow$	Cases↓	Max↓	Cases↓	Max↓	$\Delta T \uparrow$	$\Delta U \uparrow$	$\Delta S \uparrow$
UCRS (ours)	Shared	$d_T > 0,\ d_U \geq 0$	4	+0.083	+0.128	+0.073	**1**	**0.05**	**1**	**0.04**	+0.450	-1.006	-0.260
Transfer-only	Shared	$d_T > 0$	7	**+0.150**	-0.083	-0.044	3	1.10	3	0.65	**+0.517**	-1.217	-0.378
Unique-only	Shared	$d_U > 0$	5	+0.061	+0.150	+0.096	**1**	**0.05**	**1**	**0.04**	+0.428	-0.983	-0.238
Validation-only	Unique	$d_k > 0$	6	-0.394	**+0.778**	**+0.193**	3	0.75	3	0.38	-0.028	**-0.356**	**-0.140**
Screened WD	Shared	Never	0	0.000	0.000	0.000	0	0.00	0	0.00	+0.367	-1.133	-0.333
Default WD	Fixed	Never	0	-0.367	+1.133	+0.333	0	0.00	0	0.00	0.000	0.000	0.000

Bold/underline: best/second among the first four rules (ties included), not significance. d_T, d_U, d_k: MI-minus-WD selection margins (transfer/unique LP/unique kNN). Screen: early kNN code; MI: selections among nine dependent conditions. $T/U/S$: backbone/code/student evaluation accuracy. Cases counts negative contrasts; Max is the largest loss (pp), not a risk bound. Validation/default may change strength; zero WD losses do not imply global safety. Default-relative gains subtract the fixed default-WD mean, not each rule's own reference.

Task-level robustness is a separate check. Figure 7 resolves the suite reversal by task and seed: added tasks decline in every pair, while Food-101 improves throughout. Leave-one-pair-out averages retain these signs; selection is not rerun.

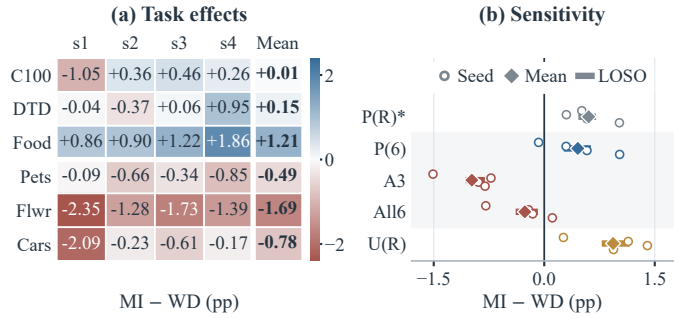

Figure 7: **Task-level effects and source-separated sensitivity.** Four pairs. (a) Six-task margins; (b) pair differences and LOSO ranges, not CIs. P/U: primary/unique; R/6: registry/six-task; A3/All6: added-three/all-six. *: distinct seed-1 records. C100/Food/Flwr: CIFAR-100/Food-101/Flowers. Selection is not rerun; ranges describe sensitivity, not complete-procedure uncertainty.

Decision stability is not held-out safety. At fixed screened strength, let $(t, u) = (\Delta_{\text{tr}}, \Delta_{\text{und}})$. The infimum ℓ_∞ perturbation changing the composition rule is

$$\rho(t,u) = \begin{cases} \min(t,u), & t > 0,\ u \geq 0, \\ \max\{(-t)_+, (-u)_+\}, & \text{otherwise}, \end{cases} \qquad (x)_+ = \max(x, 0). \tag{9}$$

Either acceptance boundary can flip a retained arm; rejection changes only after all negative margins are corrected. This radius is a sensitivity measure, not a confidence bound. H307 retains a student loss and H308 rejects a student gain: decision stability and matching do not certify safe reuse.

Scope. Evidence covers image-only factorized SSL, with a limited frozen Janus-1B check, not full UMM/world-model validation. The 20-shot and training-regime comparisons use different readouts and do not identify causal support or plasticity effects. Student comparisons use separate cohorts within this scope.

5 Conclusion

Component readouts supply selection information that backbone transfer alone misses. UCRS uses this information to optimize transfer under a calibrated code constraint. Crossed guards and prediction-preserving interface changes support matching the readout to the consumer across two students. The resulting backbone–student tradeoff depends on the intended use.

Limitations. The retrospective controls have two to four seed pairs per setting, readout-specific cohorts, and no binding unique-code constraint in the six settings. Held-out reuse covers four tasks; crossed-guard evidence spans one pretraining family. Suite reversals, false rejection and retained student losses limit generality and preclude loss-free reuse guarantees.

Reproducibility statement

The appendix documents supporting protocols and analyses. An anonymized code snapshot is provided as supplementary material and at `https://anonymous.4open.science/r/ID20229`.

AI use statement

AI tools assisted with literature exploration, manuscript preparation, mathematical exposition, and analysis code. The authors checked the resulting content and take responsibility for the accuracy, originality, attribution, and validity of the paper.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. URL `https://arxiv.org/abs/2110.01052`.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, 2018.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Imanol G. Estepa, Jesús M. Rodríguez-de Vera, Bhalaji Nagarajan, and Petia Radeva. Learning from semantic dictionaries: Discriminative codebook contrastive learning for unified visual representation and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jacob R. Gardner, Matt J. Kusner, Zhixiang Xu, Kilian Q. Weinberger, and John P. Cunningham. Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 937–945, 2014. URL `https://proceedings.mlr.press/v32/gardner14.html`.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, Amar Shah, and Ryan P. Adams. Predictive entropy search for multi-objective bayesian optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 1492–1501, 2016. URL https://proceedings.mlr.press/v48/hernandez-lobatoa16.html.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pp. 240–248, 2016. URL https://proceedings.mlr.press/v51/jamieson16.html.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

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

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Efficiently controlling multiple risks with pareto testing. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/10902.

Hongxiang Li, Hongxu Chen, Chenyang Zhu, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. SPAR: Semantic-pixel self-alignment and adaptive routing for unified multimodal models. *arXiv preprint arXiv:2606.23041*, 2026.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. In *International Conference on Learning Representations*, 2017.

Renjie Lu, Xulong Zhang, Xiaoyang Qu, Shangfei Wang, and Jianzong Wang. DIVA: Harnessing the representation divergence in unified multimodal models for mutual reinforcement. *arXiv preprint arXiv:2605.25328*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *International Conference on Machine Learning*, 2019.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. *European Signal Processing Conference*, 2007.

Wei Song, Yuran Wang, Zijia Song, Yadong Li, Zenan Zhou, Long Chen, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. DualToken: Towards Unifying Visual Understanding and Generation with Dual Visual Vocabularies. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 8086–8105, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0dbc636d5667a522cbfd92e7401ab704-Paper-Conference.pdf.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and Xudong Wang. Reconstruction Alignment Improves Unified Multimodal Models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 120095–120137, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c34ddd05eb089991f06f3c5dc36836e0-Paper-Conference.pdf.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025.

Appendix

A SELECTION ALGORITHM AND ANALYTIC SCOPE

At each strength, full-budget arms use a common training-seed set: average probe repeats within seeds, then weight seeds equally. Unequal archival cohorts are descriptive applications, not complete executions of this algorithm.

Algorithm A1 UCRS: screen strength, then constrain composition

Require: Data $\mathcal{D}$, grid Λ, horizon E, fraction b, MI weight, paired seeds and fixed probes
1: **for** $\lambda \in \Lambda$ **do**
2: Train MI-OFF for bE; record shared-code validation kNN $s_b(\lambda)$
3: **end for**
4: $\lambda^\star \leftarrow \arg\max_\lambda s_b(\lambda)$; choose the smallest strength on ties
5: Train matched MI-OFF/ON for E at $\lambda^\star$; evaluate transfer T and unique-code utility U
6: **if** a required paired readout is missing **then**
7: **return** Insufficient evidence; complete evaluation before composition
8: **end if**
9: **return** WD+MI if $T_{\mathrm{ON}} - T_{\mathrm{OFF}} > 0$ and $U_{\mathrm{ON}} - U_{\mathrm{OFF}} \geq 0$; otherwise WD

WD is feasible and wins transfer ties, so the rule solves main-paper Eq. 6. It is not a significance test. With m screening runs and r full-budget pairs, $mb + 2r$ nominal full-run equivalents exclude all probe and evaluation costs. Only the ViT-B/IN-100 study validates ranking at $b = 1/5$.

A.1 PROOF OF THE SCORE–UTILITY COUNTEREXAMPLE

For main-paper Proposition 1, consider $Z_{t,a} = (X, a(tU + N))$, $a > 0$. The first coordinate reveals the validation label X. It is independent of (U, V), where $V = a(tU + N)$, and hence adds no transfer information. Since $V \mid U = u \sim \mathcal{N}(atu, a^2)$, the equal-prior log-likelihood ratio is $2tv/a$. For $t > 0$, the optimal classifier is $\mathrm{sign}(v)$ and

$$\Pr[U(tU + N) > 0] = \Pr[t + UN > 0] = \Phi(t), \qquad \mathrm{Cov}(Z_{t,a}) = \mathrm{diag}(1, a^2(1 + t^2)).$$

At $t = 0$, transfer accuracy is $1/2$. Choosing $a = (1 + t^2)^{-1/2}$ fixes the covariance at I_2, while transfer accuracy varies. With singular-value weights $p_i = \sqrt{\lambda_i}/\sum_j \sqrt{\lambda_j}$, this fixes $r_{\mathrm{eff}} = \exp(-\sum_i p_i \log p_i) = 2$ and $\kappa = \max_i p_i = 1/2$. For the interface claim, write $Z_t = (X, V_t)$ and swap to (V_t, X). An invertible permutation preserves both full-readout optima and I_2. With the second coordinate exposed, however, $X \perp U$ gives accuracy $1/2$ instead of $\Phi(t)$. This fixes the access interface, not its semantics; it proves neither learned factorization nor student behavior.

A.2 EXACT DECISION SENSITIVITY

Proposition A1 (Two-stage decision radius)**.** *If the early maximum is unique with gap g, errors smaller than $g/2$ in every early score preserve its winner. At this fixed strength, write $(t, u) = (\Delta T, \Delta U)$. The infimum ℓ_∞ distance to the opposite composition decision is*

$$\rho(t, u) = \begin{cases} \min(t, u), & t > 0,\ u \geq 0, \\ \max\{(-t)_+, (-u)_+\}, & \textit{otherwise}. \end{cases} \tag{A1}$$

Proof. Score differences change by less than g, so the strict winner remains. Inside $\{t > 0, u \geq 0\}$, the nearest boundary has distance $\min(t, u)$. Outside, every negative coordinate must be corrected, giving the maximum deficit. Transfer strictness affects attainment, not the infimum. A zero radius allows arbitrarily small changes. Arm errors bounded by ϵ give contrast errors bounded by 2ϵ, not ϵ. The radii in Tables A5 and A10 are deterministic, not empirical error bounds or confidence intervals.

Scalar-threshold comparison. Three distinct transfer increments per environment define four constant acceptance sets, hence $4^3 = 64$ policies. Equal-condition means separate: $\max_{\boldsymbol{\tau}} \bar{Y} = \frac{1}{3}\sum_e \max_{\tau_e} \bar{Y}_e$. Accepting only H302/H308 maximizes both code and student means in this family. UCRS exceeds these maxima by $40/900$ and $22/900$ pp, respectively; mixtures cannot exceed a coordinate maximum. This evaluation-oracle enumeration is not a calibrated baseline or a maximum-transfer claim.

Reading the radius. Figure A1 separates the MI acceptance region from distance to a decision change. A large radius is not high utility; points deep inside a rejected region can also have a large radius. Strength is held fixed; the map does not estimate screening or probe-error probabilities.

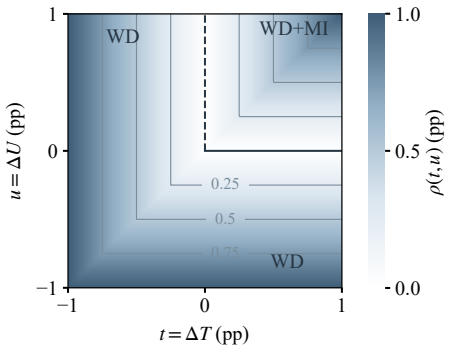

Figure A1: **Exact composition radius.** Eq. A1, not uncertainty. Dashed $t = 0$: WD; solid $u = 0$: MI if $t > 0$.

B Implementation and evaluation protocols

Architecture and supervision. Randomly initialized ViT-Ti/S/B/L trunks have 6M/22M/86M/300M parameters, patch size 16, 224×224 inputs and a class token (Dosovitskiy et al., 2021). Generation and understanding streams mask 75% and 15% of patches. Three-layer Gated-MLP encoders use GELU, an input sigmoid gate and LayerNorm, producing 256-dimensional codes. The shared encoder is reused across streams; each unique encoder serves one stream. Two-layer task heads consume concatenated shared/unique codes through $512 \to 512 \to$ output layers. The generation head predicts normalized 768-dimensional patch pixels using masked-patch MSE; the understanding head predicts one of $K = 4096$ patch-codebook indices.

Codebooks use k-means on DINOv2 14×14 patch features. The IN-1k codebook has perplexity 3908.6 (95.4% of K) with zero dead entries. This is not LEASE's adaptive centroid contrast (Estepa et al., 2026); random initialization does not imply teacher-free learning. Our within-flow CLUB adaptation likewise does not test DIVA's original cross-flow NCE-CLUB objective (Lu et al., 2026).

Optimization. AdamW uses learning rate 1.5×10^{-4}, betas $(0.9, 0.95)$, batch 256 and a one-cycle schedule with 10% warm fraction. Autocast is bf16; dependence penalties use fp32. WD spans $\{.05, .1, .2, .4, .6, .8\}$, default .05. Symmetric shared-code InfoNCE has temperature .2 and weight .5; orthogonality has weight .1; CLUB has weight .1 (gentle variant .03). These are fixed testbed settings, not outputs of screening. CLUB receives three critic steps per main step; its penalty ramps over 20 epochs after warmup. IN-100 trains for 200 epochs (100 in the horizon variant); IN-1k ViT-S/B trains for 30. Gradient checkpointing supports 40GB/24GB devices.

Training-history comparisons. P0 freezes trunk weights and mask token for 200 epochs of head training; P0-FT unfreezes for 40 further epochs. They are descriptive histories, not reproductions of DIVA's full recipe. Janus uses a frozen Janus-Pro-1B trunk with factorization heads at MI weights $.01/.03/.1$. Most comparisons have 2–3 training seeds; the selected ViT-B/IN-100 comparison has four pairs. The registry contains 94 encoders in 68 configurations. Readout-specific missingness is retained.

At uniform 20-shot support, six-task margins are -1.58/-0.55/-0.81 for Ti/S/B; all eight pair differences are negative. Frozen-regime code differences $-1.25/-1.87/-1.91$ (mean -1.68) use requested 100-shot support, not frozen-trunk 20-shot evaluation. These histories are not a causal plasticity or support trajectory.

Readouts and data separation. Shared-code in-domain kNN screens WD. In-domain diagnostic accuracy instead uses $k = 20$ kNN on normalized pooled unique-understanding codes; the IN-1k gallery has 100 images/class. For `factor_off`, the understanding output is only a baseline reference because no unique encoder exists. Transfer linear probes use frozen-trunk CLS+mean-patch features; code utility probes the pooled unique-understanding encoder. Three probe seeds are averaged within each training seed. The primary suite is CIFAR-100, DTD and Food-101; the six-task suite adds Pets,

Flowers and Cars. Support requests 100 images/class (20 in the support study), capped by availability: saved 100-shot logs contain 2,820 DTD and 2,040 Flowers training examples. The primary suite participates in composition, so it is not an independent final test. The four-task reuse comparison is in Table A4.

Diagnostics, not constraints. SVD uses a fixed stratified 1,024-image OOD pool (seed 0), with effective rank and leading singular-value share $\kappa = \sigma_1 / \sum_i \sigma_i$. Post-hoc dependence uses linear cross-correlation and normalized RBF-HSIC between shared and unique codes on the held-out split. Spectra, in-domain accuracy and broader-suite sensitivity do not enter the retention rule. Different readouts do not establish statistical equivalence or isolate domain shift under an identical representation/classifier.

Supporting materials. The anonymous repository provides training and evaluation code, per-seed readouts and analysis scripts. Early-probe records support the strength ordering; spectral arrays and cached features support the geometry diagnostics. Tables retain each readout's seed count and task aggregation. Appendices C and E specify the comparison cohorts and task-level analyses used to interpret these results.

C Measured settings and reference accounting

The complete table preserves selected and default references, readout-specific cohorts, descriptive controls and unfavorable settings. UCRS and transfer-only make identical choices in all six available-arm settings; these are the same checkpoints, not measured incremental benefits of the constraint.

Cohorts and intervals. Stored intervals summarize contrasts of arm means. Paired analyses instead use only seed-matched observations and retain the stated task weights. ViT-L transfer has OFF 2 / ON 3 available runs, not three pairs; ViT-B/IN-1k has two pairs and no stored interval. Table entries report available-seed SDs. Stored arm-mean intervals and the paired task-level sensitivity ranges describe different statistical objects.

Screening evidence. ViT-B's 40-epoch shared scores at WD $.05/.2/.4$ are 16.84/17.28/17.64; their ordering agrees with final transfer from logged epoch 19. ViT-S/IN-1k instead ranks .05 above .1 early ($23.80/23.61$), but below it in the full-budget seed-1 six-task comparison ($65.70/65.81$). Its chosen .1 is therefore not an early-screening success. Neither study establishes a universal budget fraction. All eight observed IN-1k six-suite grid differences are negative even though primary signs vary; this does not isolate a causal data-scale effect.

D Dependence controls and diagnostic limits

Matching the WD environment is essential: at $.2/.4$, transfer reaches 48.24/49.49 (+1.04/+2.29 above the matched default), while in-domain accuracy is $19.04/19.05$ versus 18.80. These are sweep contrasts, not subtraction from the multi-seed registry.

Among critic-free controls, nHSIC most reduces measured dependence but has the lowest transfer; CLUB gives the strongest transfer/unique-code means without the largest dependence reduction. This does not rule out other MI roles: estimators, optimization and regularization effects differ.

Which component fails? CS-QMI's generation-code effective rank falls from 114.5 to 3.8 without comparable trunk-transfer degradation. Separately, default CLUB changes mean trunk rank from 521.4 to 462.4 and concentration from .017 to .021 across three seeds. At WD $.4 \rightarrow .8$, the fixed seed-1 image pool gives trunk rank $537.3 \rightarrow 443.6$ and positive image-mean patch-cosine changes on all 1,024 inputs (median $+.190$). These fixed-input geometric comparisons are not independent transfer tests.

Feature-level anatomy. Figure A3 complements these aggregate readouts with cached generation-code similarities and spectra. Its generation-code ranks ($146.5 \rightarrow 6.0$) use a different feature pool/readout from the training telemetry above; their magnitudes must not be interchanged. The separate NCE trunk comparison is cross-seed and descriptive, not a matched causal control for CS-QMI.

Table A1: **Regularization and UCRS choices across six settings.** Mean, SD and within-block Δ (MI−WD), in pp; blue/brick = positive/negative. Gold UCRS (ours) choices use primary transfer and unique-code utility, applied retrospectively to available arms, not independently validated end-to-end runs. Default λ=0.05; selected 0.4 (IN-100) / 0.1 (IN-1k).

		Transfer: 3-suite			Transfer: 6-suite			$\mathcal{A}_{\mathrm{id}}$			$\mathcal{A}_{\mathrm{und}}$			
Configuration	λ	Mean	SD	Δ	Mean	SD	Δ	Mean	SD	Δ	Mean	SD	Δ	n
ViT-B / IN-100							UCRS: retain MI							
Default WD	.05	46.78	0.22	—	49.21	0.39	—	18.84	0.05	—	37.97	0.21	—	3
Default WD+MI	.05	47.93	0.46	+1.15	49.57	0.24	+0.35	18.46	0.32	−0.38	39.46	0.17	+1.49	3
WD at λ^*	.4	49.51	0.16	—	52.27	0.28	—	19.08	0.07	—	35.36	0.55	—	4
UCRS: WD+MI	.4	50.11¶	0.15	+0.60¶	52.01	0.42	−0.27	19.10	0.03	+0.03	36.30	0.42	+0.94	4
ViT-Ti / IN-100							UCRS: keep WD							
UCRS: WD	.4	40.41	0.09	—	41.79	0.12	—	17.09	0.16	—	36.03§	—	—	2
WD+MI at λ^*	.4	39.21	0.20	−1.20	40.07	0.04	−1.72	16.72	0.02	−0.38	35.35§	—	−0.68	2
ViT-S / IN-100							UCRS: retain MI (small, uncertain)							
Default WD	.05	43.33	0.53	—	45.36	0.54	—	18.49	0.15	—	36.27	0.22	—	3
Default WD+MI	.05	43.34	0.55	+0.00	45.13	0.62	−0.23	18.37	0.16	−0.13	36.81	0.73	+0.55	3
WD at λ^*	.4	45.70	0.18	—	48.09	0.14	—	18.36	0.23	—	34.47	0.65	—	3
UCRS: WD+MI	.4	45.79	0.15	+0.09	47.72	0.22	−0.37	18.53	0.06	+0.17	36.14*	0.60	+1.67	3
ViT-L / IN-100							UCRS: keep WD							
Default WD	.05	49.84	0.14	—	52.06	0.11	—	19.19	0.02	—	40.63	0.64*	—	2
Default WD+MI‡	.05	49.13	0.62	−0.71	50.50	0.37	−1.56	19.01	0.05	−0.18	40.88	0.33*	+0.25	3
UCRS: WD	.4	51.15†	0.01	—	53.64†	0.07	—	19.33	0.06	—	38.51	0.24	—	3
WD+MI at λ^*	.4	50.40	0.31	−0.74	51.53	0.18	−2.11	18.95	0.09	−0.38	39.05	0.55	+0.54	3
ViT-S / IN-1k, 30 epochs							UCRS: keep WD (transfer declines); kNN gain							
Default WD	.05	65.20	0.69	—	65.10	0.85	—	47.36	0.13	—	60.67	1.05	—	2
Default WD+MI	.05	65.21	0.86	+0.02	64.68	0.98	−0.42	47.69	0.25	+0.33	61.24	0.96	+0.58	2
UCRS: WD	.1	65.51	0.34	—	65.25	0.58	—	47.23	0.15	—	60.80	0.55	—	3
WD+MI at λ^*	.1	65.42	0.41	−0.09	64.79	0.28	−0.46	47.72	0.21	+0.48	61.40	0.67	+0.60	3
Transplant (IN-100)	.4	56.83	—		57.63	—		—	—		49.01	—		1
ViT-B / IN-1k, 30 epochs							UCRS: retain MI (n=2)							
WD at λ^*	.1	72.15	0.01	—	71.89	0.16	—	54.05	0.52	—	65.52	0.06	—	2
UCRS: WD+MI	.1	72.31	0.06	+0.17	71.61	0.14	−0.28	54.24	0.27	+0.20	67.46	0.12	+1.94	2
ViT-B selected-strength controls (descriptive)														
No factorization	.4	49.34	0.18		51.84	0.40		19.15	0.08		39.51	0.01		2
xcorr substitution	.4	49.35	0.05		51.94	0.34		18.63	0.18		33.86	0.07		3
100-epoch budget	.4	49.42	—		51.90	—		18.80	—		—	—		1

SD is over available training seeds; n applies except for the marked readout-specific cohorts. Default-strength B/S six-suite means use all six named per-dataset records. ¶For selected B ON seed 1, the registry reports 50.00 versus 49.42 for the six-task snapshot's primary subset; the four-pair margins are +0.60 and +0.46. These record sets are analyzed separately (Appendix E); task-level contrasts use the six-task records throughout. ‡w_{mi}=0.03; *unique uses 2 seeds; †transfer uses 2 seeds; §unique uses 1 seed and has no SD. Deltas use unrounded arm means; S unique (OFF 3 / ON 2) and L transfer (OFF 2 / ON 3) are not paired contrasts. Only B/IN-100 tests early screening against full-budget ordering; other rows apply composition at the reported strengths (main-paper §3.2). Stored intervals are in Appendix C. Gray rows are descriptive controls; MI is CLUB at weight 0.1 unless stated.

Neither aggregate dependence nor code-pair reductions order utility.

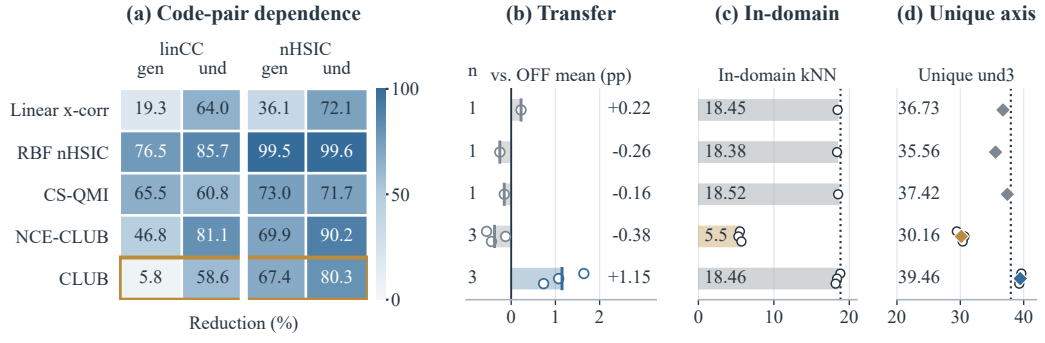

Figure A2: **Dependence, transfer, and head utility need not agree.** ViT-B, default strength. **(a)** Code-pair reductions against seed-1 OFF; nHSIC is self-measured; gold marks CLUB. **(b–d)** Bars/diamonds are arm means; open points are task seeds (n). Transfer subtracts the three-seed OFF mean (46.78%); dotted in-domain/unique references are 18.84%/37.97%. Unique-code means use a focused scale.

NCE-CLUB reduces both dependence measures more than CLUB but damages in-domain and unique-code accuracy.

Table A2: **Dependence-penalty controls at default strength.** ViT-B, WD 0.05. Penalty weight w and task training-seed count n are explicit; accuracy and SD in %, Δ in pp. Bold marks the best mean/reduction among these controls, not significance.

	Protocol		Dep. red. (%)		Transfer $\mathcal{A}_{\text{tr}}$			ID	Unique $\mathcal{A}_{\text{und}}$		
Penalty	w	n	linCC	nHSIC	Mean	SD	Δ	$\mathcal{A}_{\text{id}}$	Mean	SD	Δ
Linear x-corr	1.0	1	41.7	54.1	<u>47.00</u>	—	+0.22	18.45	36.73	—	−1.24
RBF nHSIC	1.0	1	**81.1**	**99.6***	46.52	—	−0.26	18.38	35.56	—	−2.41
CS-QMI	1.0	1	63.1	72.3	46.62	—	−0.16	**18.52**†	<u>37.42</u>	—	−0.55
NCE-CLUB	0.1	3	<u>64.0</u>	<u>80.1</u>	46.40	0.22	−0.38	5.5‡	30.16	0.63	−7.81
CLUB	0.1	3	32.2	73.9	**47.93**	0.46	+1.15	<u>18.46</u>	**39.46**	0.17	+1.49

Dependence reductions average the two code-pair reductions from one seed-1 checkpoint per arm. SD is across training seeds; — denotes an undefined SD for $n{=}1$. Δ uses common three-seed OFF means (transfer 46.78, unique 37.97), not paired contrasts. Weights differ across penalty families. †CS-QMI collapses generation-code r_{eff} (114.5→3.8). ‡NCE-CLUB's head-facing decline coexists with near-baseline trunk transfer. *nHSIC self-measures its penalty. Underlining marks second-best means/reductions; neither marking compares SDs.

The cached feature maps and spectra locate changes within individual components, rather than ranking them by aggregate dependence.

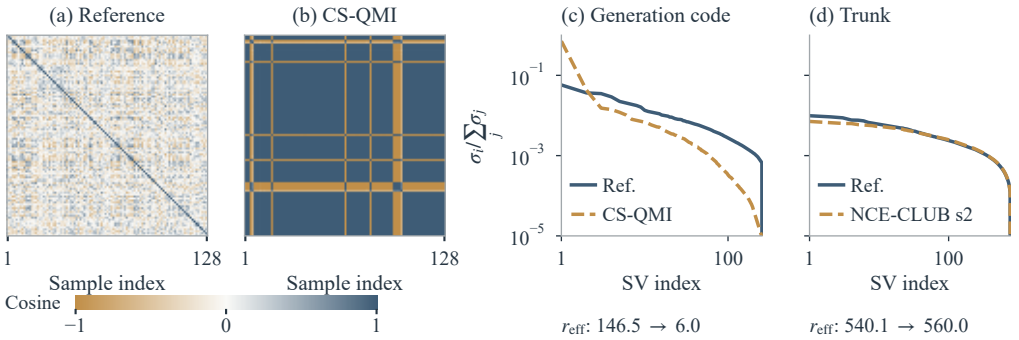

Figure A3: **Component-specific feature geometry.** (a,b) Centered generation-code cosine on the same 128 fixed examples (four per class), with a shared $[-1, 1]$ scale. (c,d) Normalized singular-value spectra use all 1,024 cached examples in each readout pool. Panel d compares reference and NCE-CLUB seed 2; it is not the trunk counterpart of the CS-QMI intervention in panel c.

From real inputs to patch geometry. Figure A4 shows one preselected input per source. For image i and patch j, we display $c_{ij} = h_{ij}^{\top}\bar{h}_i/(\|h_{ij}\|_2\|\bar{h}_i\|_2)$, where $\bar{h}_i = \frac{1}{196}\sum_{j=1}^{196} h_{ij}$. The 14×14 grid preserves token positions: darker cells mean stronger alignment with the image's mean token, not greater task relevance.

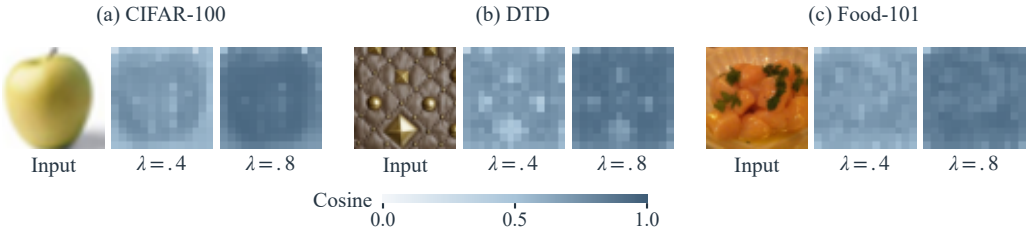

Figure A4: **Patch-token geometry on real inputs.** Fixed inputs (indices 0, 341, 682) and WD .4 / .8 maps from seed-1 ViT-B/IN-100, sharing the $[0, 1]$ scale. These feature similarities are not attention, saliency, or segmentation maps.

Inputs were fixed before inference. They complement the 1,024-image summary; this diagnostic pool is not a held-out performance test or cross-seed evidence.

Timing and limited interventions. The MI penalty activates after warmup; pre-warmup dynamics cannot characterize the full intervention. Generation-code kNN is label-dependent, not generation quality. Two lower-strength fine-tuning attempts reach 46.69 and 46.55, versus 46.64 initially: neither rescues transfer, but this does not establish irreversibility under every intervention. The full strength and timing records remain in the accompanying source ledger.

E TASK-LEVEL ROBUSTNESS AND EVIDENCE BOUNDARIES

The selected ViT-B comparison below fixes WD .4 and four pairs from one six-task snapshot. It does not pool distinct task/support protocols.

Table A3: **All four paired seeds, resolved by task.** Selected ViT-B, WD 0.4; same six-task snapshots, requested 100 examples/class, probe-repeat means. Accuracies in %; Δ (MI−WD) in pp. Mean/SD use four training seeds; SD_Δ describes paired differences.

	WD		WD+MI		Paired Δ (MI−WD)				Summary	
Task	Mean	SD	Mean	SD	Seed 1	Seed 2	Seed 3	Seed 4	Mean	SD_Δ
Primary suite										
CIFAR-100	55.77	0.21	55.77	0.55	−1.05	+0.36	+0.46	+0.26	+0.01	0.71
DTD	51.96	0.44	52.11	0.55	−0.04	−0.37	+0.06	+0.95	+0.15	0.56
Food-101	40.79	0.29	42.00	0.25	+0.86	+0.90	+1.22	+1.86	+1.21	0.46
Additional tasks in the broader suite										
Pets	62.09	0.62	61.61	0.39	−0.09	−0.66	−0.34	−0.85	−0.49	0.34
Flowers	81.41	0.27	79.72	0.57	−2.35	−1.28	−1.73	−1.39	−1.69	0.48
Cars	21.61	0.86	20.84	1.00	−2.09	−0.23	−0.61	−0.17	−0.78	0.90

These same-source cells yield +0.4558 on the first three tasks, −0.9825 on the added three, and −0.2633 on all six. The registry primary is a separate +0.6025 estimate using a different ON seed-1 record (Appendix E); the added-task effects here are computed directly from the six-task records. Differences precede display rounding.

All added tasks decline in every pair, whereas Food-101 always improves. Leave-one-pair-out means retain positive primary/unique and negative added-three/all-six signs; selection is not rerun. Figure 7 visualizes these contrasts in the main paper; the table here retains their arm-level means, SDs and paired differences without repeating the graphic.

Comparison scope. The primary registry and six-task records are analyzed separately: their primary margins are $+.6025/+.4558$ pp, with positive pair counts $4/4$ and $3/4$, respectively. ON seed 1 is 50.00 in the primary registry and 49.4167 in the six-task records. The task-level table and added-three contrast use the six-task records throughout; the primary-registry result is reported alongside it rather than pooled into a single estimate.

Configuration reuse, not constraint benefit. Four matched ViT-B pairs and three probes per task/arm/seed use at most 100 labels/class on four tasks outside both earlier suites. Average probes within seeds, then seeds and tasks equally. UCRS and transfer-only select identical checkpoints; their reuse outcomes cannot distinguish the rules. Caltech-101 remains an unfavorable case.

Table A4: **Held-out transfer.** Accuracy (%); Δ is UCRS−WD (pp).

Task	WD	**UCRS (ours)**	Δ
CIFAR-10	94.6	95.1	+0.5
SUN397	48.3	48.9	+0.6
Caltech-101	88.4	88.1	-0.3
EuroSAT	97.2	97.8	+0.6
Mean difference			**+0.35**

F COMPLETE CANDIDATE INPUTS

Selection and evaluation aggregates are listed separately. Grids are shared across three prescribed MI weights. UCRS/transfer-only/unique-only screen shared kNN; validation-only screens unique kNN. Evaluation values and unrequested full-budget strengths are unavailable to selectors.

Table A5: **Complete screening grids.** Full-budget columns are diagnostic, not early inputs.

		Early kNN		Full budget		Rank		Screen radius	
Env.	λ	Shared	Unique	Shared kNN	Transfer	Early	Final	Shared	Unique
E1	0.05	62.80	37.40	65.55	46.78	3	3	0.33	0.65
E1	0.2	64.20	36.10	65.68	49.21	2	2	0.33	0.65
E1	0.4	64.85	35.90	65.70	49.51	1	1	0.33	0.65
E2	0.05	60.50	33.40	62.30	43.33	3	3	0.13	0.25
E2	0.2	61.30	32.90	62.38	45.70	1	1	0.13	0.25
E2	0.4	61.05	31.80	62.35	45.55	2	2	0.13	0.25
E3	0.05	65.10	47.36	67.20	65.20	1	2	0.01	0.10
E3	0.1	65.08	47.56	67.23	65.26	2	1	0.01	0.10
E3	0.4	64.10	46.86	66.10	64.60	3	3	0.01	0.10

E1/E2/E3: ViT-B/IN-100, ViT-S/IN-100, ViT-S/IN-1k. Rank 1 is best. Radius is half the early top-two gap (pp), repeated within each grid; it is deterministic, not an error bar.

Identical candidate IDs share evaluations. Nine choices are dependent, not nine training repetitions. E3's ordering failure is retained.

Table A6: **All 24 candidates.** Selection inputs and independent-evaluation means; Student uses code targets.

				Selection			Evaluation		
Candidate	Env.	λ	β	Transfer	Unique LP	Unique kNN	Transfer	Unique LP	Student
C1S0	E1	0.4	0	49.51	35.36	35.60	82.10	35.40	75.20
C1S03	E1	0.4	0.03	49.76	35.81	35.59	82.30	35.75	75.40
C1S10	E1	0.4	0.1	50.11	36.30	35.62	82.45	36.10	75.60
C1S30	E1	0.4	0.3	49.86	34.56	35.65	82.35	34.30	74.55
C1V0	E1	0.05	0	46.78	37.97	37.40	81.45	37.90	75.85
C1V03	E1	0.05	0.03	46.98	38.03	37.42	81.55	37.95	75.90
C1V10	E1	0.05	0.1	46.48	37.50	37.37	81.20	37.40	75.65
C1V30	E1	0.05	0.3	46.28	36.80	37.45	81.00	36.60	75.35
C2S0	E2	0.2	0	45.70	31.25	32.90	78.80	31.20	71.40
C2S03	E2	0.2	0.03	45.85	31.50	32.89	78.90	31.35	71.50
C2S10	E2	0.2	0.1	46.00	30.70	32.94	79.00	30.35	70.95
C2S30	E2	0.2	0.3	45.60	31.50	32.98	78.60	31.40	71.60
C2V0	E2	0.05	0	43.33	32.05	33.40	78.35	32.10	71.75
C2V03	E2	0.05	0.03	43.43	32.15	33.42	78.40	32.25	71.85
C2V10	E2	0.05	0.1	43.38	31.60	33.44	78.15	31.50	71.50
C2V30	E2	0.05	0.3	43.13	30.90	33.37	78.05	30.90	71.20
C3S0	E3	0.05	0	65.20	60.67	47.36	80.60	60.67	78.10
C3S03	E3	0.05	0.03	65.40	60.77	47.37	80.70	60.62	78.06
C3S10	E3	0.05	0.1	65.45	60.59	47.36	80.75	60.72	78.14
C3S30	E3	0.05	0.3	65.15	60.22	47.38	80.50	60.32	77.95
C3V0	E3	0.1	0	65.26	60.37	47.56	80.70	60.37	78.00
C3V03	E3	0.1	0.03	65.35	60.42	47.55	80.75	60.42	78.03
C3V10	E3	0.1	0.1	65.50	60.25	47.59	80.90	60.22	77.92
C3V30	E3	0.1	0.3	65.10	59.92	47.57	80.45	59.92	77.72

$\beta = 0$: WD; LP: unique-code linear probe. The Cartesian MI menu is prescribed, not evaluation-tuned. Reused WD IDs share values. Student averages four tasks equally under the fixed recipe and never enters selection. These aggregates are not reconstructed seeds or per-task observations.

G COMPLETE-SELECTION DESIGN AND REPRODUCTION PROTOCOL

Independent units and reproduction scope. The registered block design and a reference reproduction recipe are described below, with the selection and evaluation stages treated as one procedure. Each independent outer block restarts every early grid, selects strength, fits the full-budget OFF/ON candidates and freezes every rule's configuration before opening independent evaluation cohorts. Rules share candidate fits, task weights and WD references; the three MI conditions share their environment's screen. Selection and evaluation training cohorts, probe fitting and tuning data are separate. The four named evaluation tasks are known; independence concerns the sealed cohort, not previously unseen tasks. The legacy four-pair within-selection design is distinct from independent complete-selection blocks. Neither tasks, MI conditions, teacher-nested students nor support draws increase the outer sample size. For a new replication, choose its count using a separate blinded pilot and an application-defined precision requirement; retain failed jobs, abstentions and missing readouts.

Student and distillation. Use a randomly initialized ViT-Ti/16 with a 192-dimensional mean-patch feature and a learned affine $192 \to 256$ projector. The frozen teacher supplies its mean-pooled unique-understanding code; the matched trunk target is defined in Appendix I. Disable teacher masking and dropout. Teacher and student receive the same 224-pixel crop: random resized crop with scale $[0.2, 1]$, aspect ratio $[3/4, 4/3]$, and horizontal flip probability $1/2$. With $\nu(v) = v/\max(\|v\|_2, 10^{-6})$, minimize

$$\mathbb{E}_x \left\|\nu(P_\theta f_\theta(x)) - \mathrm{sg}[\nu(z^c(x))]\right\|_2^2 .$$

Use only the environment's unlabeled training partition; evaluation-task images and labels cannot enter distillation. Keep data, initialization distribution, normalization, augmentation and optimization matched across teacher arms and targets. The reference optimizer is AdamW with learning rate 10^{-3}, betas $(0.9, 0.999)$, weight decay 0.05, batch 256 and 100 epochs: linear warmup for 10 epochs, then cosine decay to zero. No early stopping or teacher-dependent tuning is allowed. Apply this reference recipe identically to every compared teacher and target.

Student probes and aggregation. The primary frozen feature is the unprojected 192-dimensional student feature; projector-output evaluation is a separately named diagnostic. Evaluate CIFAR-10, SUN397, Caltech-101 and EuroSAT with equal task weights. For each task, use three registered support draws of up to 100 labeled training images/class, capped by availability. Standardize using that support only; fit a multinomial logistic probe with L2 inverse regularization $C = 1$, LBFGS, maximum 1000 iterations and tolerance 10^{-6}. Queries are disjoint from supports and never tune the recipe. Nonconvergence is a protocol failure, not a reason for silent retuning. Average tasks within runs, student replicates within teachers, and teacher outcomes within outer blocks. Paired uncertainty must retain all nested dependencies when resampling complete blocks. The registered block design and statistical analysis are in Appendix J.

Traceability and discriminating outcomes. Export dataset/version and image-membership hashes, teacher/codebook identities, encoder tap/pooling/masking definitions, checkpoint hashes, selection/evaluation block IDs, support/query IDs, configuration/code revisions, convergence flags and actual cost. Compare UCRS, transfer-only, unique-only, validation-only, screened/default WD and the common-threshold path on paired transfer, code and student outcomes, worst-condition loss, negative-loss sums, false rejections, retained losses and selection stability. Prespecify the reference and acceptable losses before unsealing evaluation. If simpler conservatism meets the need equally well, prefer that policy and narrow the claim. Keep each replication's execution records and outcomes together, with the same identifiers linking selection inputs, frozen configurations and evaluation readouts.

Screening readout scale. Registered shared-kNN scores near 60–65 and historical scores near 17 belong to separate evaluation records. Selection uses within-environment orderings, not the absolute difference between these scales. Cross-record comparisons require matched taps, pooling, galleries, labels, normalization, temperature, splits and checkpoints.

H FIXED-TEACHER INTERFACE CROSSING AND SENSITIVITY

Fixed interface intervention. Fix $q = (u + s)/\sqrt{2}, r = (-u + s)/\sqrt{2}$ on raw 256-dimensional codes at matching token positions, preserving paired-code information and covariance eigenvalues. Inverse mixing before original heads preserves predictions. Mean pooling commutes with mixing; per-target normalization does not preserve joint invariants. G_q replaces only UCRS's probe with a probe on q; trunk, screen, candidates, WD, supports, ties and missingness rules stay fixed. Both require $d_T > 0$ and their probe non-decrease. No angle search; identical choices are uninformative.

Students and independence. Cross u/q targets with ViT-Ti and random ResNet18 (He et al., 2016): standard batch normalization, a 512-dimensional global-average feature, affine $512 \to 256$ projector and pre-projector probe. Use Appendix G's recipe, without target-specific tuning or equal-FLOP claims. Each outer block repeats grids and selection before independent evaluation. Candidate/architecture/target results are reused across guards; two student initializations and three supports stay nested.

Table A7: **Fixed-teacher interface disagreements.** Student columns give same-strength MI-minus-WD evaluation gains (pp).

Case	d_T	d_U	d_Q	ViT dS_u	ViT dS_q	ResNet dS_u	ResNet dS_q
H301	+0.25	+0.45	+0.10	+0.20	+0.10	+0.15	+0.12
H302	+0.60	+0.94	-0.20	+0.40	-0.20	+0.35	-0.25
H303	+0.35	-0.80	+0.35	-0.65	+0.35	-0.60	+0.40
H304	+0.15	+0.25	-0.15	+0.10	-0.15	+0.12	-0.12
H305	+0.30	-0.55	+0.25	-0.45	+0.35	-0.50	+0.32
H306	-0.10	+0.25	+0.20	+0.20	+0.20	+0.18	+0.16
H307	+0.20	+0.10	-0.12	-0.04	-0.15	-0.06	-0.20
H308	+0.25	-0.08	+0.10	+0.04	+0.10	+0.07	+0.08
H309	-0.05	-0.45	-0.35	-0.15	-0.10	-0.20	-0.12

d_T, d_U, d_Q are same-strength selection contrasts (pp); d_T, d_U and ViT-u agree with main-paper Table 2. Shared candidates and WD references are reused; nine conditions are not independent repeats. The data registry contains all 24 candidate readouts.

Complete-selection precision. For main-paper Table 4's four directional contrasts (ViT-u/q, ResNet-u/q), the registered summaries contain $n = 24$ complete-block contrasts with sample covariance $\Sigma = .0096 I_4 + .0048\mathbf{1}\mathbf{1}^\top$ pp^2. For these four contrasts and two within-student sums, use $\bar{d} \pm t_{23,1-.05/12}\sqrt{a^\top \Sigma a/24}$ under independent, approximately normal outer contrasts. Directional intervals are $[.098, .240]/[.074, .215]$ for ViT and $[.089, .231]/[.082, .223]$ for ResNet. This family is separate from Appendix J's four and two comparisons, not joint twelve-contrast coverage. The rotation alone does not imply this precision; inner student and support repeats do not increase n. Probe and student execution costs are not included in these intervals.

Original scalar sensitivity retained. Assume independent uniform $[-a, a]$ errors across distinct arms, stages and metrics, with half-widths $(s, T, U, k) = (.01, .05, .05, .01)$ or $(.05, .15, .15, .03)$ pp. Reuse candidate-ID/WD draws across rules; evaluation means stay fixed. Maximize exact expected $T + wS$ over all real per-environment transfer thresholds shared across MI weights/strengths, named rules and shared/unique screens crossed with T-only/U-only/UCRS gates. Noisy T+A/G_q are excluded; there is no held-out loss guarantee. Only the tabulated w range favors UCRS.

Table A8: **Scalar-preference envelope.** Conditional model summaries.

	Supported w		UCRS outcomes			
Noise	Lower	Upper	$\mathbb{E}T$	$\mathbb{E}S$	L_S (%)	Gap†
Narrow	0.56278	1.00000	80.584	74.973	11.11	0.004387
Wider	0.50143	0.78275	80.587	74.969	6.18	0.003429

†Gap at $w = .6$ for $J_w = \mathbb{E}T + w\mathbb{E}S$; ties included. Gaps against the strongest competitor are weighted points, not confidence bounds. L_S averages accepted-MI student losses over all nine decisions using each draw's local WD; main-paper Table 5 instead fixes shared-screen WD. No empirical precision, cost or application weight is inferred.

I Component-matched crossed comparison

Matched controls. The original 24 candidates, shared screen and same-strength WD remain fixed. T+A accepts MI iff $\Delta T > 0$ and $\Delta A \geq 0$. A probes a 256-dimensional frozen-trunk Gaussian projection with orthonormalized columns, fixed without labels before outcomes and shared across arms within each environment/block. Code/trunk probes share supports, fitting, missing-readout and tie conventions. Dimension matching does not match information content.

For MI indicators a_i^U, a_i^A and same-strength MI-minus-WD student gains d_i^C, d_i^A, shared candidate evaluations give exactly

$$\Omega = \frac{1}{9}\sum_i (a_i^U - a_i^A)(d_i^C - d_i^A) = \frac{.90 + .40 + .75 + .16 + .04}{9} = .250.$$

The nonzero terms are H303/H304/H305/H307/H308. H307 contributes .16 despite code/trunk losses $.04/.20$: interaction is not safety. This aggregate identity does not determine the separately reported block statistics. Both students also see raw images (Appendix G); matching concerns target usefulness, not isolated component information.

Table A9: **Trunk readouts for the 24 candidates.** Projection-probe and trunk-target student accuracies, with same-strength contrasts.

Candidate	A	S_A	ΔA	ΔS_A	Candidate	A	S_A	ΔA	ΔS_A
C1S0	45.00	75.00	0.00	0.00	C2V0	41.50	71.30	0.00	0.00
C1S03	45.20	75.18	+0.20	+0.18	C2V03	41.58	71.40	+0.08	+0.10
C1S10	45.55	75.35	+0.55	+0.35	C2V10	41.38	71.15	-0.12	-0.15
C1S30	45.10	75.25	+0.10	+0.25	C2V30	41.30	71.00	-0.20	-0.30
C1V0	44.00	75.60	0.00	0.00	C3S0	58.00	77.80	0.00	0.00
C1V03	44.12	75.72	+0.12	+0.12	C3S03	57.88	77.60	-0.12	-0.20
C1V10	43.80	75.45	-0.20	-0.15	C3S10	58.08	77.88	+0.08	+0.08
C1V30	43.50	75.20	-0.50	-0.40	C3S30	57.80	77.68	-0.20	-0.12
C2S0	42.00	71.00	0.00	0.00	C3V0	58.10	77.70	0.00	0.00
C2S03	41.85	70.70	-0.15	-0.30	C3V03	58.15	77.76	+0.05	+0.06
C2S10	42.25	71.30	+0.25	+0.30	C3V10	57.98	77.62	-0.12	-0.08
C2S30	42.05	70.90	+0.05	-0.10	C3V30	57.90	77.50	-0.20	-0.20

Candidate IDs, all original selection/evaluation values and code-target student means are unchanged from Table A6. A, S_A are percentages; Δ columns subtract the $\beta = 0$ candidate at the same environment/strength (pp). The fixed projection has 256 output dimensions. Shared candidate IDs reuse identical evaluations; no new source seeds or per-task observations are assigned.

Preserved counterexamples. H308 falsely rejects a useful arm. Unique-only also meets the illustrative requirements: student gain .140 versus UCRS's .118, with identical maximum backbone sacrifice .25 and local student loss .04. Appendix J quantifies this tradeoff, not dominance.

Table A10: **Reference-relative utility and distinct failure events.** Utility readings are evaluation percentages; contrasts and composition radii are points.

		Evaluation unique LP			UCRS contrast			Event indicator		
Case	UCRS ID	Local WD	Default	UCRS	Local	Default	ρ	Bind	FR	RL
H301	C1S03	35.40	37.90	35.75	+0.35	−2.15	0.25	0	0	0
H302	C1S10	35.40	37.90	36.10	+0.70	−1.80	0.60	0	0	0
H303	C1S0	35.40	37.90	35.40	0.00	−2.50	0.80	1	0	0
H304	C2S03	31.20	32.10	31.35	+0.15	−0.75	0.15	0	0	0
H305	C2S0	31.20	32.10	31.20	0.00	−0.90	0.55	1	0	0
H306	C2S0	31.20	32.10	31.20	0.00	−0.90	0.10	0	0	0
H307	C3S03	60.67	60.67	60.62	−0.05	−0.05	0.10	0	0	1
H308	C3S0	60.67	60.67	60.67	0.00	0.00	0.08	1	1	0
H309	C3S0	60.67	60.67	60.67	0.00	0.00	0.45	0	0	0

Local WD shares the UCRS-selected strength; default WD fixes the environment's declared default. Bind: UCRS differs from transfer-only; FR: a rejected MI arm improves evaluation unique utility; RL: retained MI loses evaluation unique utility against local WD. FR and RL are distinct. ρ is the infimum ℓ_∞ contrast perturbation changing composition, not empirical noise or a confidence interval.

J Complete-selection precision and application limits

Block design and registered statistics. The registered analysis uses $n = 24$ complete-selection blocks, each restarting all three environment grids. Use one matched teacher pair per environment/MI condition, sharing WD across three MI weights, then two student initializations per target and three support draws per student. Aggregate within block; these inner repeats do not increase n. The registered complete-selection means agree with the aggregate contrasts. For $D^g_{c,b} = S_{c,b}(\pi_g) - S_{c,b}(\pi_T)$, $g \in \{U, A\}$, $c \in \{C, A\}$, the interaction is

$$\Omega_b = (D^U_{C,b} - D^U_{A,b}) - (D^A_{C,b} - D^A_{A,b}).$$

In order $(D^U_C, D^U_A, D^A_C, D^A_A)$, the registered sample covariance is

$$\Sigma = \begin{pmatrix} .0144 & .0056 & .0076 & .0056 \\ .0056 & .0100 & .0044 & .0032 \\ .0076 & .0044 & .0121 & .0044 \\ .0056 & .0032 & .0044 & .0144 \end{pmatrix} \quad \text{(pp squared)}.$$

These block summaries are distinct from fixed-teacher aggregate arithmetic and the bounded-noise sensitivity model. For contrast vector a, use $s^2_d = a^\top \Sigma a$ and $\bar{d} \pm t_{23,1-.05/8} s_d/\sqrt{24}$: independent, approximately normal outer contrasts and a four-comparison Bonferroni correction are working assumptions. Positive definiteness alone does not establish coverage.

For the direct UCRS/unique-only comparison, the registered analysis uses the same $n = 24$ complete-block design and reports means for $(T_{\text{UCRS}} - T_{\text{U-only}},\ S_{C,\text{U-only}} - S_{C,\text{UCRS}})$. Its sample covariance is $\left(\begin{smallmatrix} .0009 & .0003 \\ .0003 & .0009 \end{smallmatrix}\right)$ pp^2. This separate secondary family uses $t_{23,1-.05/4}$; no joint six-contrast confidence claim is made.
The primary intervals have half-widths below .10 pp for interaction and .08 for the other contrasts. These are summary-statistic intervals, not simultaneous guarantees across all conditions or downstream consumers.

Table A11: **Complete-selection precision.** Paired contrasts across $n = 24$ outer blocks.

Contrast (pp)	Mean	SD	Lower	Upper
S_C: UCRS−TO	+0.118	0.120	+0.051	+0.184
S_C: UCRS−(T+A)	+0.124	0.106	+0.066	+0.183
S_A: (T+A)−UCRS	+0.126	0.134	+0.051	+0.200
Guard×target	+0.250	0.171	+0.155	+0.345
Separate secondary family (UO: unique-only)				
T: UCRS−UO	+0.022	0.030	+0.008	+0.037
S_C: UO−UCRS	+0.022	0.030	+0.008	+0.037

Paired Student-t intervals use Bonferroni 95% separately within the four primary and two secondary contrasts, not jointly over six. Nested repeats do not increase n.

Illustrative requirements and assumed costs. For an accepted calibrated backbone and code-target student, require gain ≥ 0.05 pp over TO, maximum condition-wise backbone sacrifice ≤ 0.30 pp versus TO, and student loss ≤ 0.10 pp versus local WD. Anchors give $0.118/0.25/0.04$ pp; unique-only also meets these requirements. Its better student mean costs backbone accuracy: a tradeoff, not universal UCRS superiority. These illustrative thresholds are not evidence of preregistered deployment requirements. Mean intervals are not simultaneous condition-wise loss bounds; default WD better preserves default-relative student utility.

Per environment/block, screening and distinct teacher fits cost $3(0.2) + 1 + 3 = 4.6$ nominal full-run equivalents. Assume 0.30 for transfer readouts and 0.06 extra per auxiliary probe: TO costs 4.90, UCRS/T+A 4.96. The increment is 20% of TO readouts (illustrative ceiling 25%), or 1.22% of the restricted selector total. These are not GPU timings and exclude independent evaluation teachers, two-target students, failed runs and hardware/cache effects. The observed interaction therefore does not establish deployment value: adopting UCRS additionally requires an acceptable backbone–student tradeoff and measured evaluation overhead for the intended use.