THE REPRESENTATION-CONTINUATION GAP: LOW-RANK STATES VS. LOW-RANK DYNAMICS IN VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

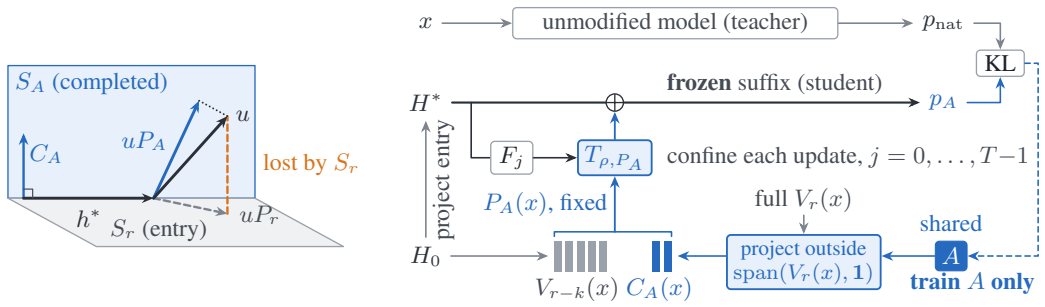
Low-rank activations make a Vision Transformer look compressible, but compression has to survive the remaining residual computation: after a state has been represented in a small subspace, the later blocks must still compute there. We show that these two requirements can diverge. A subspace can reconstruct the current state, or entry, nearly exactly while failing to support the suffix computation, a failure we call the *representation-continuation gap*. In eight frozen ViT settings spanning supervised, self-supervised, and CLIP pretraining, confining only the last two blocks to an almost lossless entry space reduces accuracy by 1 to 67 points (median 12). The gap persists under distribution shift and at mid depth, and it appears even though late-layer subspaces are image-specific and stable across views. The obstruction is directional rather than merely low-rank: at a 64-direction budget, replacing eight entry directions with directions read from the input's natural future raises CLIP-FT ViT-B from 3.9% to 82.5%, while random directions reach 4.9%. We explain this with two facts: a fixed continuation space must contain directions created after the entry, and whether an omitted direction matters depends on how strongly the output responds to it. We then introduce *functional subspace completion* (FSC), a small self-distillation of subspace directions. FSC keeps the dominant entry directions, learns a shared source of complementary directions through the frozen suffix by matching the original model's predictions, and uses no future states at test time. With only 6,144 learned parameters per ViT-B classifier (0.007% of the frozen model), FSC raises accuracy on 5,000 held-out images from 50.6% to 70.5% on AugReg ViT-B and from 4.0% to 77.6% on CLIP-FT ViT-B relative to the entry basis, against 83.0% and 84.5% for unconstrained continuation. FSC improves on the entry basis in all eight settings and transfers to shifted test sets without retraining. Low-rank spaces should therefore be evaluated not only by the states they reconstruct, but by the computation they allow to continue. Code is available in the supplementary material.

1 INTRODUCTION

Low-rank activations make wide transformers look as if they contain much smaller computations. In Vision Transformers (ViTs) (Dosovitskiy et al., 2021), the token features of a single image can be highly compressible: in one ViT with 384 channels, 61 singular directions recover 95% of last-layer feature energy for about 99% of images (Tian et al., 2026). This structure motivates many compression methods for vision and language transformers (Yu & Wu, 2023; Ashkboos et al., 2024; Yuan et al., 2023; Wang et al., 2025; Chowdhury et al., 2026; Shihab et al., 2026). Yet a residual transformer is not a snapshot: later attention and MLP branches can write channel directions absent from the state they receive. This paper asks: **when a subspace represents the current ViT state nearly losslessly, can the remaining layers continue computing inside that same subspace?** If so, a compressed suffix could run at reduced width; if not, reconstruction describes a stored state rather than a runnable computation. Appendix G counts the multiply-adds of such a folded suffix.

For one input, call the state where a restriction begins the *entry* and the following layers the *suffix*. A natural entry subspace is the span of the entry's leading right singular directions. *Representation sufficiency* asks whether projection leaves the entry nearly unchanged. *Continuation sufficiency* asks

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.



(a) Same entry, different support

(b) FSC in the frozen suffix

Figure 1: **Entry reconstruction and continuation support can diverge. (a)** Equal-rank spaces $S_r = \text{span}(V_r)$ and $S_A = \text{span}(V_{r-k}, C_A)$ both preserve h^*, but retain different components of a later update u (orange: the component S_r discards); C_A lies outside $\text{span}(V_r, \mathbf{1})$. **(b)** FSC fixes an input-dependent projector throughout the suffix and applies T_{ρ,P_A} to each residual update. Natural-teacher KL trains only the shared source A; test-time selection uses the current entry, not future states.

whether the output survives when every suffix attention and MLP update is confined to the same subspace (Figure 1(a)). This stronger condition is what a suffix compressed into that space would need: the space must carry residual dynamics, not merely describe the state those dynamics produced.

We test continuation by intervention: hold the input, frozen weights, readout, and entry fixed, and change only the directions available to later updates (Geiger et al., 2025; Zhang & Nanda, 2024). This exposes a *representation-continuation gap*. In a complete-entry control, the subspace is the full row space of the entry's 197 tokens; the entry changes by at most 6.88×10^{-7} relative error, and free continuation matches the natural model. Yet confining only the last two blocks to that space reduces accuracy by 15.7 points on AugReg ViT-B, 33.2 on CLIP-FT ViT-B, and 1 to 67 points (median 12) across eight frozen encoder-readout settings (Figure 3). The gap persists under five shifts, appears at mid depth, and grows with suffix length. Late-layer subspaces are image-specific and stable across views (Figure 2), so they are meaningful objects; they are just not necessarily runnable state spaces.

The missing support is small but specific. At a 64-direction budget, replacing eight entry directions with directions from the same input's natural future raises accuracy from 52.8% to 79.6% on AugReg ViT-B and from 3.9% to 82.5% on CLIP-FT ViT-B, close to unconstrained continuation; eight random directions reach only 54.7% and 4.9% (Table 1). This privileged future arm is a diagnostic, not a test-time method: the entry basis fails because the budget is spent on the wrong directions.

Our theory explains what those directions must do. A fixed space preserves the residual trajectory only if it contains every later update; the extra dimension is the rank of those updates outside the entry row space (Proposition 1). Prediction preservation can need fewer directions, but not arbitrary ones: discarded updates reach logits through suffix Jacobians, which can amplify a small update and suppress a large one, so output sensitivity determines which omissions matter (Proposition 2). Thus a low-rank space should complement the entry in directions that protect predictions.

Functional subspace completion (FSC, Figure 1(b)) implements this criterion. It keeps the $r - k$ dominant entry directions and obtains the remaining k from a source matrix shared across images. For each input, the source is projected away from the current entry support, yielding an input-dependent complement fixed throughout the suffix. We train only this source, through the frozen suffix, to match the unmodified model's predictions. This is self-distillation of subspace selection: teacher and student share backbone and readout, but the student suffix is constrained. At test time FSC uses the entry and learned source, not future states.

On 5,000 held-out images at 64 directions, FSC raises accuracy from 50.6% to 70.5% on AugReg ViT-B and from 4.0% to 77.6% on CLIP-FT ViT-B relative to the entry basis, against 83.0% and 84.5% for unconstrained continuation from the same entry (Table 2). The learned source has only $768 \times 8 = 6{,}144$ trainable scalars for a ViT-B classifier, about 0.007% of the frozen model. FSC improves on the entry basis in all eight settings, at budgets 32 and 128, under raw projection, and

under shift without retraining. Over four blocks, larger learned complements shrink the remaining gap, as expected when a longer suffix creates more required directions.

Our contributions are threefold. First, we define and test *continuation sufficiency*, showing by controlled intervention that a nearly lossless representation space can still fail the next residual blocks. Second, we give a structural explanation: continuation requires directions outside the entry's row space, and their importance is set by output sensitivity. Third, we introduce FSC, a lightweight self-distilled completion of an entry subspace that improves every setting tested and closes most of the two-block gap in seven of eight. Together, these results make residual dynamics a design criterion for low-rank spaces: a useful subspace must be judged not only by what it reconstructs, but by what it lets the network continue to compute.

2 Related work

Our work connects representation analysis and compression. Transformer activations often have lower-dimensional structure than their channel width suggests; we ask the stricter question of whether such a space is also a valid state space for the residual computation that remains.

Representation geometry and rank. Representation-similarity methods compare features across layers, architectures, examples, and views (Raghu et al., 2021; Kornblith et al., 2019; Chen et al., 2020; Wang & Isola, 2020); ViT rank studies show that per-image token features can be highly compressible, while a single common encoding can fail across inputs (Tian et al., 2026). We ask a within-input continuation question: do later residual branches write directions absent from the entry?

Efficient transformers and functional compression. Efficient transformer methods reduce attention matrices or tokens (Wang et al., 2020; Rao et al., 2021; Bolya et al., 2023), while channel-space compression exploits low-rank activations or activation-calibrated weights through feature compression, principal-component slicing, activation-aware factorizations, and sensitivity-weighted reconstruction (Yu & Wu, 2023; Ashkboos et al., 2024; Yuan et al., 2023; Wang et al., 2025; Chowdhury et al., 2026; Shihab et al., 2026). These methods usually choose directions from existing states or weights. Our diagnostic exposes a reconstruction-blind failure mode by holding the input, entry, budget, backbone, and readout fixed and changing only directions available to later residual updates.

Learning low-rank interventions. LoRA learns low-rank weight updates, and ReFT learns low-rank representation interventions in frozen language models (Hu et al., 2022; Wu et al., 2024a). FSC also learns through a frozen network, but the learned object is a complement to the entry space: the directions of the model's *own* later updates that remain available under a fixed channel budget.

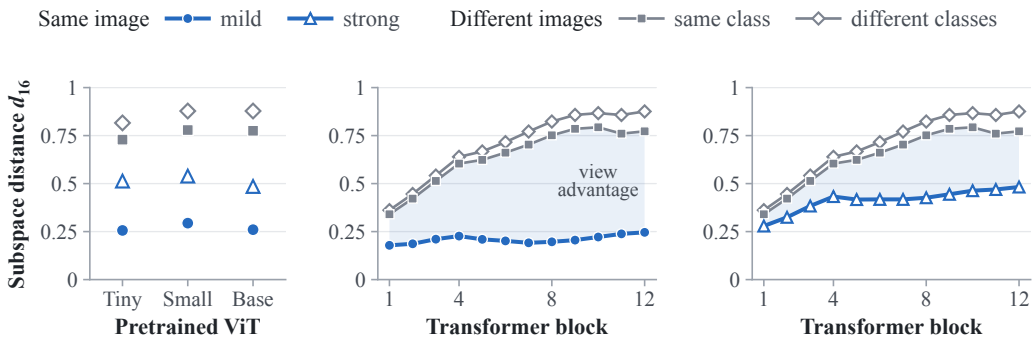
Reduced dynamics and causal interpretation. Projection-based model reduction separates snapshot approximation from fidelity to later dynamics and the input-output map (Rowley, 2005; Benner et al., 2015). Activation-patching and causal-abstraction work similarly show that claims about representations depend on the intervention (Geiger et al., 2025; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Makelov et al., 2024; Wu et al., 2024b). We use the same intervention view to test whether a low-rank ViT representation space is sufficient for the computation that follows.

3 Low-rank states and constrained continuation

We fix three objects used throughout: an *entry* state, a per-image channel subspace, and a controlled suffix rollout. Each residual branch is still evaluated by the frozen model; only its update, immediately before addition, is restricted to a chosen channel space. The prefix, weights, and readout are unchanged, so any loss comes from this restriction.

From a represented state to a runnable suffix. For one input, let $H_j \in \mathbb{R}^{N\times D}$ be the token-by-channel state at residual site j. Site 0 is the entry, and sites $1, \dots, T$ are the attention and MLP residual additions in the suffix. If F_j denotes the original branch, including its original normalization, the natural and constrained rollouts under an intervention rule ρ are

$$H_{j+1} = H_j + F_j(H_j), \qquad \widetilde{H}_{j+1} = \widetilde{H}_j + T_{\rho,P}\Big(F_j(\widetilde{H}_j)\Big). \tag{1}$$



(a) Final block (b) Mild views (c) Strong views

Figure 2: **Late-layer subspaces are image-specific yet stable across views.** Rank-16 squared chordal distances between patch-token subspaces. **(a)** Final-block means for three pretrained widths (10,000 images). **(b,c)** Mild and strong views across AugReg ViT-B depth (200 images). Shading marks the view advantage a_ℓ: different same-class images minus two views of one image.

Here $P = VV^\top$ is an input-specific channel projector with $V \in \mathbb{R}^{D\times r}$ orthonormal, reused for every suffix update. Raw projection sets $T_{\mathrm{raw},P}(U) = UP$. Our primary *token-norm* rule restores each update token's channel mean and centered norm (Appendix B.1), changing mainly direction while preserving scalar scale. Its image lies in $\mathrm{span}(V, \mathbf{1})$, so r is a *nominal* budget: it counts selected columns of V, with the all-ones mean direction $\mathbf{1}$ restored separately. We report cross-entropy (CE), accuracy, and KL divergence from the natural model.

Candidate entry spaces come from the state itself: unless noted, $V_r(x)$ contains the leading r right singular directions of image x's entry, including the class token. This favorable, per-input reconstruction space is then tested as a carrier for residual additions that have not yet occurred.

Per-image subspaces are stable enough to be worth testing. We first check that per-image subspaces are not arbitrary SVD artifacts. Let $V_{\ell,r}(x)$ contain the leading r right singular directions of image x at block ℓ, and compare subspaces by normalized squared chordal distance

$$d_{\ell,r}(x, x') = 1 - \|V_{\ell,r}(x)^\top V_{\ell,r}(x')\|_F^2/r, \tag{2}$$

which is zero for identical subspaces and one for orthogonal ones. At $r = 16$, two mild views of one image have final-block distances $0.256/0.294/0.260$ in pretrained ViT-Tiny/Small/Base, versus $0.729/0.779/0.775$ for two same-class images (Figure 2(a)). In AugReg ViT-B (Steiner et al., 2022), same-image views stay aligned with depth while different images separate (Figure 2(b,c)); pretraining sharpens this profile relative to random initializations (Appendices B.3 and B.4).

This geometry shows that the subspaces are image-specific, view-stable, and shaped by training.

4 Why continuation is harder than representation

A low-rank entry space must both represent the current state and serve as a state space for later residual computation. The first does not imply the second, because a branch can write a channel direction absent from its input. For a fixed raw projector, exact path preservation requires the entry and every later update to lie in the projector's range.

Proposition 1 (Fixed-space closure and the complementary dimension)**.** *Let $U_j = F_j(H_j)$ be the natural residual updates, and suppose $\widetilde{H}_0 = H_0 = H_0P$. Under raw projection, the constrained rollout equals the natural trajectory at every site if and only if $U_jP = U_j$ for every j. Consequently, the minimum fixed-space dimension for exact trajectory preservation is*

$$r_{\mathrm{path}} = \mathrm{rank}\begin{bmatrix} H_0 \\ U_0 \\ \vdots \\ U_{T-1} \end{bmatrix} = \mathrm{rank}\begin{bmatrix} H_0 \\ H_1 \\ \vdots \\ H_T \end{bmatrix}. \tag{3}$$

Writing P_0 for the projector onto the row space of H_0 and $r_0 = \mathrm{rank}(H_0)$,

$$r_{\mathrm{path}} - r_0 = \mathrm{rank} \begin{bmatrix} U_0(I - P_0) \\ \vdots \\ U_{T-1}(I - P_0) \end{bmatrix}. \tag{4}$$

Proof sketch. Containment of every natural update gives the natural trajectory by induction, and equality of consecutive states forces containment. States and successive differences span the same rows; decomposing updates into entry-space and orthogonal components gives Equation (4) (Appendix A).

Equation (4) is invisible to entry reconstruction: it counts post-entry directions outside the entry row space. Stable view geometry does not remove it; even a perfectly view-consistent rank-one entry can rotate into a different next-state direction, so a fixed entry projector erases the computation (Appendix A). The issue is closure, not subspace noise.

Prediction fidelity can need less than exact path preservation: some missing directions are logit-null, while small discarded components can be amplified. As in reduced dynamics, the question is which defects reach the readout (Rowley, 2005; Benner et al., 2015); Appendix A.3 gives a norm bound.

Proposition 2 (Output-sensitive continuation defects)**.** *Fix P, let $E_0 = H_0 - \widetilde{H}_0$, and write $R_j(H) = F_j(H) - T_{\rho,P}(F_j(H))$. Introduce an intervention-strength parameter ϵ through*

$$H_0^\epsilon = H_0 - \epsilon E_0, \qquad H_{j+1}^\epsilon = H_j^\epsilon + F_j(H_j^\epsilon) - \epsilon R_j(H_j^\epsilon). \tag{5}$$

Assume the branches, intervention maps, and logit readout are twice continuously differentiable in a neighborhood of the natural trajectory. Let J_{in} be the natural suffix logit Jacobian with respect to the entry and J_j the corresponding Jacobian with respect to an additive state perturbation immediately after update j. Define

$$q_P = J_{\mathrm{in}} \,\mathrm{vec}(E_0) + \sum_{j=0}^{T-1} J_j \,\mathrm{vec}(R_j(H_j)). \tag{6}$$

For logits z^ϵ, set $p^\epsilon = \mathrm{softmax}(z^\epsilon)$ and $p = p^0$. Then

$$z^\epsilon = z^0 - \epsilon q_P + O(\epsilon^2), \qquad D_{\mathrm{KL}}(p \| p^\epsilon) = \frac{\epsilon^2}{2} q_P^\top [\mathrm{diag}(p) - pp^\top] q_P + o(\epsilon^2). \tag{7}$$

Equation (7) gives a functional, not geometric, criterion. Suffix Jacobians and the KL metric decide whether omissions reinforce, cancel, or pass through nearly null directions (Martens, 2020); update energy is only a proxy. FSC follows this criterion by choosing the complement that preserves predictions, optimizing teacher KL through the full-strength constrained suffix.

5 Interventions reveal the representation-continuation gap

We test whether an almost lossless entry is enough, and whether a few complementary directions matter at fixed entry and rank.

Models and protocol. We evaluate eight frozen encoder-readout settings: AugReg ViT-B, AugReg2 ViT-B, CLIP-FT ViT-B, AugReg ViT-S/L, DINO-B with a linear readout (Caron et al., 2021), MAE-B with its frozen readout (He et al., 2022), and zero-shot CLIP-B (Radford et al., 2021). Unless stated otherwise, the entry precedes the last two blocks and the suffix contains their four attention/MLP residual additions over 197 tokens. Diagnostics use 2,000 class-balanced ImageNet-validation images (Russakovsky et al., 2015); significance means a 95% bootstrap interval, simultaneous unless noted, excludes zero (Appendices B, D, and E.3). Table 32 lists every arm.

5.1 A nearly lossless entry can still be a poor continuation space

The *complete-entry* control removes entry information loss as a confound. Both arms first project the entry onto all 197 columns of its thin right-singular basis, changing the state by at most 6.88×10^{-7} relative error under token-norm. The *entry-only* arm then continues freely and matches natural

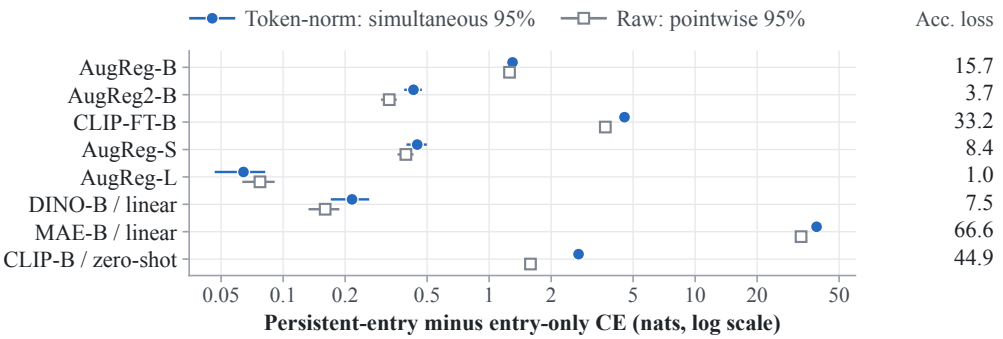


Figure 3: **A lossless entry basis can still fail as a continuation space.** CE increase when four later updates are also restricted to the complete 197-column entry basis (2,000 images, two-block suffix). The right column gives token-norm accuracy loss in points. Bars are 95% bootstrap intervals, simultaneous for token-norm and pointwise for raw projection; the axis is logarithmic.

accuracy; the *persistent-entry* arm restricts the four later updates to the same basis, measuring directions needed during continuation.

The difference is large (Figure 3): accuracy falls by 15.7 points on AugReg-B, 33.2 on CLIP-FT, 44.9 on zero-shot CLIP-B, and 66.6 on MAE-B; even AugReg-L loses 1.0 point. CE rises significantly in all eight settings, including raw projection. A basis containing essentially the entire entry can still be a poor two-block state space.

A future basis pinpoints the missing support. For each input, we compute an equally large basis from the natural entry and the next two block outputs. It represents the entry *less* accurately (4%–12% mean relative error), but tracks the residual path. Outside MAE-B, projecting the entry onto it (*entry-only-future*) and also constraining the four updates (*persistent-future*) cost at most 0.8 and 1.6 accuracy points, respectively, versus 1.0 to 44.9 for the complete entry basis; in MAE-B the summed cost is 25.4 rather than 66.6 (Table 8). The worse representation basis is the better continuation space.

The gap is not an endpoint artifact. We repeat the complete-entry control on ImageNetV2, ImageNet-Sketch, ImageNet-R, ImageNet-A, and ImageNet-C (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021a;b; Hendrycks & Dietterich, 2019). Persistent-entry is furthest from the model's own output in all settings and datasets, significantly in all 70 comparisons (Table 13). AugReg-B loses 15.5–19.3 points under shift, matching its 15.7 in-distribution loss, while persistent-future loses at most 1.6. Longer and earlier suffixes show the same pattern: four constrained AugReg-L blocks raise loss from 1.0 to 4.1 points (Table 16); at mid depth, two blocks cost 1.4–68.1 points, while persistent-future stays within 1.4 except in MAE-B (Table 17).

5.2 At fixed budget, the decisive object is the complement

We next ask whether a small budget can work if a few columns are spent differently. Every arm starts from the same projected entry $H^* = T_{\text{token_norm},P_{56}}(H_0)$, with P_{56} built from the natural entry, and restricted arms use 64 columns. Entry64 uses V_{64}; completion arms keep V_{56} and add eight directions outside $\text{span}(V_{64}, \mathbf{1})$, so gains must come from directions unavailable to Entry64. The *future* arm obtains them from natural future states; the random arm samples them (Appendix C).

Only eight columns change the outcome (Table 1). Future-derived directions raise accuracy from 52.8% to 79.6% on AugReg-B and from 3.9% to 82.5% on CLIP-FT, within five points of free continuation; random directions leave the same models at 54.7% and 4.9%. CE falls relative to both Entry64 and random complements in all eight settings, significantly in all sixteen comparisons (Appendix D); under four shifted datasets it also falls in all three tested settings (Table 14). Useful continuation directions exist at the same entry and budget, but they are not the next entry singular vectors; with only 64 columns, the future arm even beats the complete 197-column basis in seven of eight settings (Table 8). Section 6 learns a test-time replacement for the privileged future arm.

Table 1: **Eight directions decide most of the outcome.** Accuracy (%) on 2,000 diagnostic images; common entry H^*, last two blocks, 64 token-norm columns. Random is mean $\pm$ sample SD over five seeds. Future replaces eight entry directions; the last column gives its CE reduction relative to Entry64 in nats, with a simultaneous 95% CI.

Encoder-readout	Free ↑	Entry64 ↑	Random ↑	Future ↑	CE reduction ↑ (95% CI)
AugReg-B	84.25	52.80	54.73 ± 0.62	**79.55**	$2.149\ [2.052, 2.247]$
AugReg2-B	85.40	75.30	75.80 ± 0.48	**83.15**	$0.670\ [0.609, 0.731]$
CLIP-FT-B	84.80	3.90	4.86 ± 0.22	**82.50**	$5.253\ [5.138, 5.369]$
AugReg-S	80.55	57.15	57.41 ± 0.27	**74.50**	$1.065\ [0.977, 1.153]$
AugReg-L	85.45	84.70	84.64 ± 0.07	**85.90**	$0.093\ [0.071, 0.116]$
DINO-B / linear	72.30	63.55	65.59 ± 0.21	**70.40**	$0.372\ [0.308, 0.435]$
MAE-B / linear	58.95	0.30	0.29 ± 0.14	**1.80**	$23.447\ [22.075, 24.819]$
CLIP-B / zero-shot	68.50	11.40	11.31 ± 0.38	**54.30**	$3.281\ [3.081, 3.482]$

6 Learning the continuation complement

The future arm shows that a few directions can rescue continuation, but it sees the natural future. *Functional subspace completion* (FSC) replaces that access with a tiny direction source learned once by self-distillation. FSC leaves the backbone, readout, entry, and task unchanged; it learns which directions of the frozen model's own later updates remain available under a fixed channel budget.

6.1 A shared source gives per-image complements

FSC splits a nominal budget of r directions into $r - k$ entry directions and k continuation directions. Let $V_r(x)$ be the entry basis for image x, and let $A \in \mathbb{R}^{D\times k}$ be the only trainable parameter. With $k = 8$ in the main experiments (Appendix D.1 varies k),

$$\begin{aligned} U_r(x) &= \mathrm{orth}([V_r(x), \mathbf{1}]), \\ C_A(x) &= \mathrm{orth}\big((I - U_r(x)U_r(x)^\top)A\big)\,, \\ V_A(x) &= [V_{r-k}(x), C_A(x)], \qquad P_A(x) = V_A(x)V_A(x)^\top. \end{aligned} \tag{8}$$

The projection against $U_r(x)$, the whole rank-r entry support, puts the learned columns outside the baseline entry basis. Because the cleaning is per image, one shared source A becomes an input-dependent complement $C_A(x)$, fixed throughout that suffix (Figure 1(b)).

Every continuation arm starts from the same projected entry $H^* = T_{\rho,P_{r-k}}(H_0)$; FSC changes only the later-update projector. For each calibration image, the unmodified model supplies a teacher distribution and the constrained suffix a student. We train only A:

$$\min_A \ \frac{1}{|\mathcal{D}_{\mathrm{fit}}|} \sum_{x\in\mathcal{D}_{\mathrm{fit}}} D_{\mathrm{KL}}(p_{\mathrm{natural}}(x) \,\|\, p_A(x; H^*(x)))\,. \tag{9}$$

Gradients pass through the frozen suffix, so directions are chosen by their effect on constrained computation rather than update energy or reconstruction error. Natural future activations initialize A; a random orthonormal start reaches the same accuracy (Appendix E.6). At test time, FSC uses only the entry and learned source. The loss is standard distillation (Hinton et al., 2015); what is learned is the subspace in which the unchanged model continues computing.

Training and held-out confirmation. We fit A on 6,000 images, tune on 2,000, train each configuration for ten Adam epochs at learning rate 3×10^{-3} from three initializations, and keep the epoch with lowest natural-teacher KL on a fixed 256-image tuning subset. Tables 2 and 3 report seed means and sample standard deviations. A ViT-B source has $768 \times 8 = 6{,}144$ trainable scalars, about 0.007% of the frozen 86.6M parameters (Appendix C.3). Unless noted, evaluation uses 5,000 held-out ImageNet-validation images; later settings reuse the same recipe unchanged (Appendix E).

Table 2: **Held-out confirmation on 5,000 images.** Accuracy (%), CE, and KL from the unmodified model's predictions (CE/KL in nats). All non-natural arms start from the same entry H^*; free continuation leaves the last two blocks unrestricted, while constrained arms use nominal budget $r = 64$. A_0 and FSC entries are mean ± sample standard deviation over three seeds.

	AugReg ViT-B			CLIP-FT ViT-B		
Policy	Acc.↑	CE↓	KL↓	Acc.↑	CE↓	KL↓
Natural teacher	83.98	0.587	0.000	85.04	0.639	0.000
Free from H^*	83.00	0.614	0.029	84.50	0.682	0.030
Entry basis V_{64}	50.64	2.958	2.369	4.02	6.833	5.659
Clean geometric G_8	34.38	3.470	2.865	72.88	2.615	1.773
Matched init. A_0	34.34 ± 0.19	3.470 ± 0.001	2.866 ± 0.001	72.79 ± 0.19	2.616 ± 0.009	1.774 ± 0.008
FSC (selected A_*)	**70.53** ± 0.26	**1.509** ± 0.004	**0.925** ± 0.001	**77.60** ± 0.24	**1.500** ± 0.002	**0.807** ± 0.002

Table 3: **FSC across settings and budgets.** Accuracy (%) on 5,000 confirmation images; last two blocks, token-norm, $k = 8$. A_0, FSC, and $\Delta_{\text{train}} = \text{CE}(A_0) - \text{CE}(A_*)$ (nats) are mean ± sample SD over three seeds; differences are paired by seed.

Setting	r	Entry↑	A_0↑	FSC↑	Free↑	Δ_{train}↑
Further encoder-readout settings						
AugReg2-B	64	74.72	76.93 ± 0.01	**80.09** ± 0.06	84.26	1.649 ± 0.006
AugReg-L	64	84.06	84.09 ± 0.01	**84.64** ± 0.08	84.84	0.0935 ± 0.0002
DINO-B / linear	64	62.42	66.38 ± 0.05	**71.05** ± 0.13	71.90	0.215 ± 0.001
MAE-B / linear	64	0.34	0.31 ± 0.03	**14.28** ± 0.47	57.56	9.867 ± 0.131
CLIP-B / zero-shot	64	11.28	28.31 ± 0.13	**55.53** ± 0.23	67.80	1.560 ± 0.010
Width and budget						
AugReg-S	64	56.40	61.93 ± 0.22	**71.04** ± 0.11	79.42	0.456 ± 0.002
AugReg-B	32	26.86	28.34 ± 0.08	**66.91** ± 0.17	76.10	2.356 ± 0.003
AugReg-B	128	61.38	48.31 ± 0.19	**74.24** ± 0.07	83.84	1.595 ± 0.005
CLIP-FT-B	32	2.98	44.04 ± 0.21	**66.95** ± 0.49	81.24	2.728 ± 0.016
CLIP-FT-B	128	20.34	78.37 ± 0.15	**80.46** ± 0.03	85.06	0.707 ± 0.005

6.2 Eight learned directions close most of the two-block gap

At a nominal budget of 64 directions, the entry basis reaches 50.64% on AugReg-B and 4.02% on CLIP-FT, while free continuation from the same entry reaches 83.00% and 84.50% (Table 2). Replacing eight entry directions by FSC raises accuracy to 70.53% and 77.60%, recovering 61% and 91% of the lost accuracy; teacher KL falls from 2.369 to 0.925 nats and from 5.659 to 0.807.

Clean G_8 and its slightly perturbed copy A_0, the matched initialization from which training starts, are untrained future-projector sources; they help CLIP-FT but hurt AugReg-B (about 73% versus 34% accuracy). Training through Equation (9) improves both models, lowering CE by 1.96 nats on AugReg-B (95% interval 1.90–2.02) and 1.12 on CLIP-FT (1.08–1.15), with every trained source beating its initialization (Appendices E.3 and E.7). The future-projector start is not needed: a random start, which uses no data, trains to 70.79% and 77.55%. The trained sources outperform training-free sensitivity sources, and from the same entry FSC's 64 columns beat all 197 entry directions, which reach 65.9% and 41.1%; confined outside the complete entry support, FSC retains 93% and 98% of its CE gain (Appendix E.6). On the development set, it also beats an input-conditioned future-projector predictor and approaches the privileged future arm (Appendix F). FSC is also far more accurate than removing the two blocks (8.98% and 62.82%; Appendix G). Eight shared directions leave a 12.5-point and 6.9-point gap to free continuation; 32 raise FSC to 76.0% and 78.2% (Table 22).

6.3 Same recipe across settings, budgets, and shifts

Keeping $k = 8$ and the recipe fixed, FSC beats both the entry basis and matched initialization in every row of Table 3. It lifts zero-shot CLIP-B from 11.28% to 55.53% (free: 67.80%), DINO-B

from 62.42% to 71.05% (71.90%), AugReg2-B from 74.72% to 80.09% (84.26%), and MAE-B from 0.34% to 14.28%, where training lowers CE by 9.87 nats. AugReg-L is within 0.20 points of free continuation. Budget changes preserve the effect: at $r = 128$, CLIP-FT's entry basis reaches 20.34%, FSC 80.46%, and free continuation 85.06%; raw projection also benefits (Appendix E.4).

The learned directions transfer to shifted inputs. Without retraining, using each shifted image's own entry, FSC improves over the entry basis and lowers KL relative to matched initialization in all nine model-dataset pairs (Table 23). CLIP-FT rises from 2.5% to 66.7% on ImageNetV2 (free: 74.9%), and AugReg-B from 34.2% to 49.1% on ImageNet-C (67.1%). FSC recovers 23% to 89% of the entry-basis loss, a fraction that generally falls with shift severity.

A longer suffix needs a larger complement. Over four blocks, eight learned directions no longer suffice: training raises accuracy from 2.69% to 30.41% on AugReg-B and from 13.91% to 37.04% on CLIP-FT, below free continuation at 82.98% and 85.02%. Larger complements help: 16 and 32 directions reach 40.15%/48.99% on AugReg-B and 46.41%/55.28% on CLIP-FT, shrinking the remaining gap from 52.6 to 30.6 points and from 48.0 to 28.3 (Figure 4 and Table 22). This matches Equation (4): longer suffixes can create more needed directions, but they remain learnable.

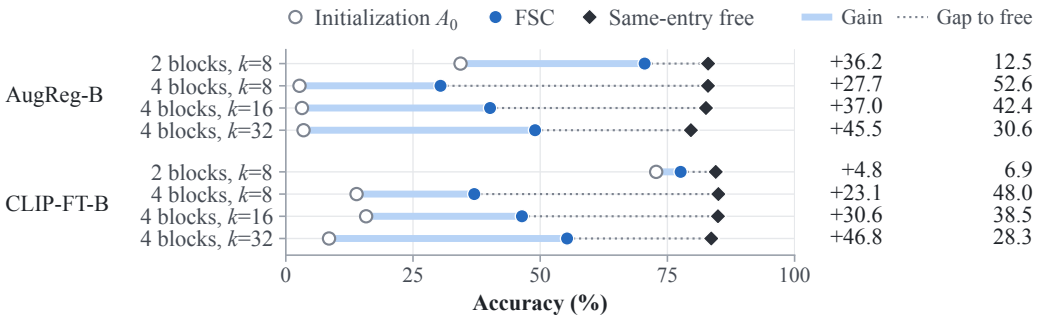


Figure 4: **Longer suffixes require larger learned complements.** Mean accuracy on the 5,000-image confirmation set (nominal $r = 64$, token-norm) as learned directions k vary and the entry keeps $64 - k$. Shaded segments join initialization to FSC; dotted segments continue to same-entry free continuation. Right columns report segment lengths in points.

7 Conclusion

A low-rank subspace can describe a ViT state almost perfectly and still be a poor state space for the computation that follows. In the complete-entry control, the entry is preserved to within 10^{-6} relative error, yet forcing the last two blocks to keep computing in that space costs up to 67 accuracy points. The gap appears across eight frozen encoder-readout settings, under supervised, self-supervised, and CLIP pretraining, on shifted inputs, and at late and middle depths.

The obstruction is structural and functional. Residual branches can write directions absent from their input, so exact continuation requires closure under later updates (Proposition 1); prediction fidelity depends on how strongly the output responds to what is omitted (Proposition 2). The missing support is often small but specific: eight privileged future directions bring CLIP-FT from 3.9% to 82.5%. FSC turns that diagnostic into a test-time rule: with an eight-column shared source, 6,144 learned parameters for a ViT-B classifier, and no future states, it raises AugReg-B from 50.6% to 70.5% and CLIP-FT from 4.0% to 77.6% at the same 64-direction budget, improves every setting tested, transfers under shift, and closes most of the two-block gap in seven of eight.

Low-rank spaces should therefore be judged by the computation they support as well as by the states they reconstruct. For compression, the same-entry free-continuation gap diagnoses whether a subspace is runnable; for representation analysis, it tests whether later layers compute within a claimed activation subspace. Because FSC finds the missing directions from the entry alone, the completed support is a concrete target for a reduced-width suffix (Appendix G); block-varying spaces as in SliceGPT (Ashkboos et al., 2024) and language-model continuation are natural next steps.

AI USE STATEMENT

Large language models were used to polish writing, assist with code drafting, and give feedback on the design of experiments. The authors reviewed all AI-assisted work and are responsible for the manuscript, proofs, references, code, and reported results.

ETHICS STATEMENT

The study uses existing ImageNet-derived data and pretrained models, with no new human-subject data collection. The original image and model licenses remain applicable, and images and pretrained weights are not redistributed. Because FSC is trained to preserve the original models' predictions, their errors and biases may also be preserved. The distribution-shift experiments report fidelity to the original model under shift rather than fairness or robustness interventions.

REPRODUCIBILITY STATEMENT

The appendices specify the proofs, interventions, datasets, frozen models and readouts, source construction, optimization, and statistical comparisons. The supplementary code implements the interventions and FSC for all eight settings. From the public checkpoints and datasets, its scripts rerun the subspace measurements of Section 3, the complete-entry and common-entry controls in every window, projection rule, and dataset, and the development-set comparisons; evaluate the supplied sources and ablations and the removed window; train a source with the recipe of Table 7 from each initialization of Appendix E.6; compute class-bootstrap intervals; and test the analytical identities on constructed systems. Appendix H.2 describes the code, and Appendix H lists the model assets and preprocessing needed to rerun inference.

REFERENCES

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vXxardq6db.

Peter Benner, Serkan Gugercin, and Karen Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531, 2015. doi: 10.1137/130932715. URL https://doi.org/10.1137/130932715.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://doi.org/10.1109/ICCV48922.2021.00951.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Md Mokarram Chowdhury, Daniel Agyei Asante, Ernie Chang, and Yang Li. IMPACT: Importance-aware activation space reconstruction. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3975–3992, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.183. URL https://aclanthology.org/2026.acl-long.183/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024. URL https://arxiv.org/abs/2404.15255.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2021a. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15257–15266, 2021b. doi: 10.1109/CVPR46437.2021.01501. URL https://doi.org/10.1109/CVPR46437.2021.01501.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uPv9Y3gmAI5.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/747d3443e319a22747fbb873e8b2f9f2-Abstract.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

C. W. Rowley. Model reduction for fluids, using balanced proper orthogonal decomposition. *International Journal of Bifurcation and Chaos*, 15(03):997–1013, 2005. doi: 10.1142/S0218127405012429. URL https://doi.org/10.1142/S0218127405012429.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. URL https://doi.org/10.1007/s11263-015-0816-y.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Beyond variance: Knowledge-aware LLM compression via Fisher-aligned subspace diagnostics. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47938–47954, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2214. URL https://aclanthology.org/2026.acl-long.2214/.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=4nPswr1KcP.

Huiyuan Tian, Bonan Xu, and Shijian Li. From per-image low-rank to encoding mismatch: Rethinking feature distillation in vision transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2Ud1nkQrVZ.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric Xing. Learning robust global representations by penalizing local predictive power. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. arXiv:2006.04768, 2020. URL https://arxiv.org/abs/2006.04768.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LNYIUouhdt.

Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962. Curran Associates, Inc., 2024a. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75008a0fba53bf13b0bb3b7bff986e0e-Abstract-Conference.html.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to Makelov et al. (2023)'s "interpretability illusion" arguments. arXiv:2401.12631, 2024b. URL https://arxiv.org/abs/2401.12631.

Hao Yu and Jianxin Wu. Compressing transformers: Features are low-rank, but weights are not! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11007–11015, 2023. URL https://doi.org/10.1609/aaai.v37i9.26304.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. ASVD: Activation-aware singular value decomposition for compressing large language models. arXiv:2312.05821, 2023. URL https://arxiv.org/abs/2312.05821.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

APPENDIX

Appendix A proves the two propositions. Appendix B gives the intervention protocol and subspace measurements, and Appendix C describes source construction and FSC training. Appendices D–F report the controlled diagnostics, held-out confirmation, extensions, ablations, and development-set source comparisons behind Sections 5 and 6. Appendix G gives operation counts for a folded suffix and compares FSC with removing the window, and Appendix H lists the implementation and reproduction assets. Boldface highlights the focal method or effect named in each table caption; it does not denote a ranking across configurations or a significance test.

A PROOFS AND CONSEQUENCES FOR CONSTRAINED COMPUTATION

The two propositions answer two questions: which space reproduces every state of a residual trajectory, and, when some update directions are omitted, which of them reach the readout. We treat them in that order.

A.1 FIXED-SPACE CLOSURE AND THE COMPLEMENTARY DIMENSION

We first characterize the fixed spaces that reproduce the natural trajectory exactly. All row spaces below are subspaces of the channel space $\mathbb{R}^D$. Let P be a fixed orthogonal projector and assume $\widetilde{H}_0 = H_0 = H_0 P$. For raw projection,

$$\widetilde{H}_{j+1} = \widetilde{H}_j + F_j(\widetilde{H}_j)P. \tag{A.1}$$

If $U_j P = U_j$ at every natural site, induction starting from $\widetilde{H}_0 = H_0$ gives $\widetilde{H}_j = H_j$ for every j. Conversely, equality of both consecutive states gives

$$H_j + U_j = H_j + U_j P, \tag{A.2}$$

so $U_j P = U_j$. Thus exact trajectory preservation is equivalent to containment of the rows of $H_0, U_0, \ldots, U_{T-1}$ in the chosen space. The smallest space with this property is their row span, proving the first expression for r_{path}.

Since $H_j = H_0 + \sum_{i<j} U_i$, each state lies in the span generated by the entry and updates. Conversely, $U_j = H_{j+1} - H_j$, so every update lies in the span generated by the states. The two spans, and hence the two ranks in Equation (3), are equal.

Let $\mathcal{S}_0 = \mathrm{row}(H_0)$, let P_0 project onto $\mathcal{S}_0$, and let $\mathcal{R}$ be the row span of the stacked $U_j(I - P_0)$. For each update, its component $U_j P_0$ lies in $\mathcal{S}_0$. Subtracting that component shows that the full trajectory span is

$$\mathcal{S}_0 + \mathcal{R} = \mathcal{S}_0 \oplus \mathcal{R}, \tag{A.3}$$

where the sum is orthogonal because $\mathcal{R} \subseteq \mathcal{S}_0^{\perp}$. Dimensions therefore add, giving Equation (4). This proves Proposition 1.

Trajectory rank and FSC's budget. The dimension r_{path} is what exact reproduction of this trajectory costs under one fixed raw projector and the original residual functions. Preserving only the output can cost less, and so can coordinates that change from layer to layer. FSC works in that cheaper regime: at a chosen budget it optimizes output fidelity, in the main setting with 56 directions that approximate the entry and eight selected outside the full $\mathrm{span}(V_{64}, \mathbf{1})$. That exclusion leaves the entry's other 133 row-space directions available, but the learned complement does not favor them and keeps 93% and 98% of its gain without them (Appendix E.6). The proposition thus explains why an entry space is not closed under the computation, and the ViT experiments measure how much of the output a small, well-chosen completion preserves.

Token count, width, and the token-norm rule. In every model we constrain, the $N = 197$ tokens are fewer than the D channels (384, 768, or 1,024). Hence $r_0 \le N < D$, and a basis can hold the entire entry while excluding at least $D - N$ channel directions, which is what makes the complete-entry control of Section 5 informative. Appendix D.2 measures how much of each natural update falls in the excluded directions. Proposition 1 is stated for raw projection. The image of

the token-norm map lies in $\mathrm{span}(V, \mathbf{1})$, so reproducing a trajectory exactly under token-norm still requires the rows of the entry and of every update to lie in that span, and the number of basis columns this takes is at least $r_{\mathrm{path}} - 1$.

A.2 Perfect view stability does not imply closure

Stable view geometry cannot guarantee closure, as a two-dimensional example shows. Use column vectors. Let $h_0(a) = ae_1$, $a > 0$, and choose an orthogonal R with $Re_1 = e_2$. Define one residual branch by $F(h) = (R - I)h$. All entry views have the same rank-one subspace, and all natural final views have the same rank-one subspace. Yet for $P = e_1 e_1^\top$,

$$h_1 = ae_2, \qquad \widetilde{h}_1 = ae_1 + P(ae_2 - ae_1) = 0. \tag{A.4}$$

The readout $e_2^\top h$ changes by a, and scaling a scales this error without changing any view-subspace distance. Each state still has a one-dimensional description, but its direction changes from one layer to the next. Stable view geometry at every layer therefore coexists with an entry space that is not closed across layers.

A.3 Residual error propagation

An imperfect entry and discarded updates add up along the suffix as follows. For raw projection, let $E_j = H_j - \widetilde{H}_j$. Adding and subtracting $F_j(H_j)P$ yields the exact identity

$$E_{j+1} = E_j + [F_j(H_j) - F_j(\widetilde{H}_j)]P + F_j(H_j)(I - P). \tag{A.5}$$

An orthogonal projector has operator norm at most one. Local L_j-Lipschitz continuity in the Frobenius norm therefore implies

$$e_{j+1} \le (1 + L_j)e_j + d_j, \quad e_j = \|E_j\|_F, \quad d_j = \|F_j(H_j)(I - P)\|_F. \tag{A.6}$$

Iterating this scalar inequality, with an empty product read as one, gives

$$\|H_T - \widetilde{H}_T\|_F \le e_0 \prod_{j=0}^{T-1}(1 + L_j) + \sum_{j=0}^{T-1} d_j \prod_{k=j+1}^{T-1}(1 + L_k). \tag{A.7}$$

The first term propagates the entry damage e_0, and the second accumulates the defects d_j of the discarded updates.

For a general map $T_{\rho,P}$, add and subtract $T_{\rho,P}(F_j(H_j))$ instead. If $T_{\rho,P}$ is locally K_j-Lipschitz on the relevant update pairs and $b_j = \|F_j(H_j) - T_{\rho,P}(F_j(H_j))\|_F$, then

$$e_{j+1} \le (1 + K_j L_j)e_j + b_j. \tag{A.8}$$

The same product bound holds with L_j replaced by $K_j L_j$ and d_j by b_j. The constants are local to the compared states and updates, and for token-norm a small projected centered norm can make K_j large. In the complete-entry control the rescaling stays modest, between 1.1 and 2.0 on average (Appendix D.2). The complete-entry control of Section 5 removes the first term of the bound, so what it measures is the effect of the second.

A.4 Output-sensitive continuation defects

The next step identifies which of these errors reach the output. Flatten each state to h_j and write $f_j(h) = h + F_j(h)$. At $\epsilon = 0$, the homotopy in Proposition 2 follows the natural trajectory. Let $v_j = \partial h_j^\epsilon / \partial \epsilon\big|_{\epsilon=0}$. Differentiating gives

$$v_0 = -\operatorname{vec}(E_0), \qquad v_{j+1} = Df_j(h_j)v_j - \operatorname{vec}(R_j(H_j)). \tag{A.9}$$

The derivative of R_j is multiplied by ϵ and therefore contributes zero at the natural trajectory. A perturbation added immediately after update j enters at state H_{j+1}, so $J_j = \partial z / \partial \operatorname{vec}(H_{j+1})$ and $J_{\mathrm{in}} = \partial z / \partial \operatorname{vec}(H_0)$, both evaluated on the natural trajectory. Unrolling this linear recursion and composing with the final logit readout gives

$$\left.\frac{\partial z^\epsilon}{\partial \epsilon}\right|_{\epsilon=0} = -J_{\mathrm{in}} \operatorname{vec}(E_0) - \sum_{j=0}^{T-1} J_j \operatorname{vec}(R_j(H_j)) = -q_P. \tag{A.10}$$

The twice-continuously-differentiable assumption supplies the logit expansion in Equation (7). Attention, the GELU MLPs, LayerNorm, the readouts used here, and raw projection are smooth. The token-norm map of Equation (B.1) is smooth wherever an update token has a nonzero centered norm and its projection has a centered norm above the clamp, and this is the condition that the assumption places on the natural trajectory.

Let $\mathcal{A}(z) = \log \sum_c e^{z_c}$. With $p = \mathrm{softmax}(z)$ and $\delta = z^\epsilon - z$,

$$\begin{aligned} D_{\mathrm{KL}}(p \| \,\mathrm{softmax}(z+\delta)) &= \mathcal{A}(z+\delta) - \mathcal{A}(z) - \nabla\mathcal{A}(z)^\top \delta \\ &= \int_0^1 (1-t)\, \delta^\top \nabla^2 \mathcal{A}(z+t\delta)\delta \, dt. \end{aligned} \tag{A.11}$$

The Hessian is $\mathrm{diag}(p) - pp^\top$ at z, giving the quadratic expansion after substituting $\delta = -\epsilon q_P + O(\epsilon^2)$. The KL expression has no linear term, so the second-order logit remainder does not alter its leading quadratic term. This proves Proposition 2.

Relation to Fisher-weighted selection. The matrix $S = \mathrm{diag}(p) - pp^\top$ is the Fisher information of the categorical output distribution with respect to its logits, so $J_j^\top S J_j$ is that Fisher information pulled back to a perturbation of the state after update j. Gradient- and Fisher-weighted compression criteria (Chowdhury et al., 2026; Shihab et al., 2026) use sensitivity of this kind at the site being compressed. The leading term of Equation (7) applies it to the *combined* defect q_P, summed over the entry and every update of the suffix before the metric acts, and the teacher KL that FSC minimizes is the same quantity at full intervention strength. Appendix E.6 builds five training-free sources on this sensitivity, and none approaches the learned source; for the entry basis itself, the leading term underestimates the measured KL by a factor of 7 to 42.

Why a scalar energy statistic is insufficient. Writing $q_P = \sum_a q_a$ to include the entry term, the leading discrepancy is

$$\tfrac{1}{2} \sum_a q_a^\top S q_a + \sum_{a<b} q_a^\top S q_b, \qquad S = \mathrm{diag}(p) - pp^\top. \tag{A.12}$$

The cross terms can be positive or negative. Even a large defect can be annihilated by a suffix Jacobian, or can produce a class-independent logit shift that the softmax ignores. Discarded Frobenius energy therefore misses both downstream sensitivity and the interactions among sites. The expansion is local in ϵ, which is why FSC optimizes the actual KL at full intervention strength.

A finite-strength directional identity. There is also an exact trajectory-dependent error decomposition. Let f_j and g_j be the natural and constrained state maps, $\delta_j = f_j(h_j) - g_j(h_j)$, and $e_j = h_j - \widetilde{h}_j$. Under absolute continuity along the relevant state segment,

$$e_{j+1} = M_j e_j + \delta_j, \qquad M_j = \int_0^1 Dg_j(\widetilde{h}_j + te_j)\, dt. \tag{A.13}$$

Unrolling yields an ordered sum of propagated defects, which tracks directional errors at finite strength. Its matrices are evaluated along the constrained trajectory.

A.5 Output stability and prediction preservation

Finally, the constrained model inherits the stability and the predictions of the natural model whenever its logit error is small. If natural logits satisfy $\|z(x) - z(t(x))\| \le \delta(x,t)$ and constrained logits obey $\|\widetilde{z}(x) - z(x)\| \le E(x)$ for each input, the triangle inequality gives

$$\|\widetilde{z}(x) - \widetilde{z}(t(x))\| \le \delta(x,t) + E(x) + E(t(x)). \tag{A.14}$$

Output stability of the natural model thus transfers to the constrained model once continuation errors are controlled, so the quantity to bound is the continuation error. If the natural top-1 margin is $m > 0$ and $\|\widetilde{z} - z\|_\infty < m/2$, the constrained model keeps the natural model's predicted class.

Constructed examples. The supplementary code checks the row-span identity, projected rollouts, the rank-one counterexample, and the local KL expansion numerically on small smooth residual systems. The ViT measurements are in Appendices D and E.

B EXPERIMENTAL DESIGN AND SUBSPACE GEOMETRY

B.1 NORMALIZATION AND INTERVENTION SITES

The primary intervention projects each update and then restores its mean and scale. For one token update u, let $q = uP$, and let $\mu(u)$ denote its channel mean. The map is

$$T_{\text{token_norm,P}}(u) = \mu(u)\mathbf{1} + \frac{\|u - \mu(u)\mathbf{1}\|_2}{\max(\|q - \mu(q)\mathbf{1}\|_2, 10^{-12})}[q - \mu(q)\mathbf{1}]. \tag{B.1}$$

The map preserves the channel mean and, when the projected update has nonzero centered norm, the centered norm. Its image lies in $\mathrm{span}(V, \mathbf{1})$ when $P = VV^\top$, and the clamp handles the zero-norm case. Appendix A.3 shows how small projected norms enter the error bound, while Appendix D.2 reports the observed rescaling factors and compares token-norm with raw projection.

Each intervention transforms a branch update immediately before residual addition, while the branch itself is evaluated at full width. For the 12-block ViTs, the two-block suffix starts after block 9 and acts at the four attention and MLP additions in blocks 10 and 11, and the four-block suffix starts after block 7 and acts at all eight additions in blocks 8 to 11 (block indices are zero-based). Each suffix has its own entry and cached states. The diagnostics use the same last-two-block construction at other depths, with an additional last-four-block condition for AugReg-L.

The complete-entry control has four arms, each with 197 basis columns. *Entry-only* projects the entry onto the entry basis and then continues freely. *Persistent-entry* also restricts every update in the window to that basis. *Persistent-future* uses an equally large basis from the unnormalized Gram sum of the input's natural entry and subsequent block outputs, and *entry-only-future* applies that future basis only at the entry. The future basis reads the input's own later states, so it is available only as a diagnostic.

Contrast A1 is persistent-entry minus entry-only, and contrast A2 is persistent-entry minus persistent-future. The A1 arms share one entry, as do the two future arms. A2 compares two complete projection policies whose entries can differ; under token-norm, the future-basis entry differs from the entry-basis entry by 4.0%–12.4% relative error on average across settings. Entry-only-future is reported descriptively, while the distribution-shift and mid-depth controls use the first three arms.

Table 4 summarizes the eight encoder-readout settings, and Appendix H lists their checkpoints.

Table 4: Diagnostic encoder-readout settings. B/S/L denote Base/Small/Large; all models use patch size 16, 224×224 inputs, and 197 tokens. Natural accuracy is measured on the 2,000 diagnostic images.

Encoder	Width	Blocks	Frozen readout	Acc. (%)↑
AugReg-B	768	12	Original 1,000-class head	85.05
AugReg2-B	768	12	Original 1,000-class head	85.35
CLIP-FT-B	768	12	ImageNet-fine-tuned 1,000-class head	86.15
AugReg-S	384	12	Original 1,000-class head	82.00
AugReg-L	1024	24	Original 1,000-class head	86.45
DINO-B / linear	768	12	Official linear readout on the final class token and mean patch features	79.25
MAE-B / linear	768	12	Frozen class-token, batch-normalization, and linear readout	67.05
CLIP-B / zero-shot	768	12	Official prompt ensemble with the checkpoint logit scale	68.40

B.2 DATASETS AND THE QUESTIONS THEY ADDRESS

Table 5 maps the main experiments to their role in the paper.

Table 5: Experimental studies and their roles. Rows group the experiments used to establish the gap, train FSC, and test extensions; sample construction is summarized below.

Study	Inputs	Role
View geometry	AugReg Tiny/Small/Base; rank-16 patch-token spaces	Measure image-specific subspaces and view stability across depth.
Random-weight view control	AugReg-B and three random initializations	Separate effects of pretraining from geometry induced by architecture and augmentations.
Complete/common-entry diagnostics	Eight encoder-readout settings; 2,000 in-distribution images	Isolate entry reconstruction from continuation support and compare direction choices at a shared entry.
Shift and window controls	Shifted ImageNet variants; mid-depth and four-block windows	Test whether the continuation gap persists off distribution and away from the final two blocks.
FSC confirmation	Two ViT-B models; 5,000 held-out images	Evaluate selected FSC sources against entry, initialization, clean geometric, and free-continuation references.
Configuration extensions	Additional models, budgets, windows, projection rules, and shifted inputs	Test the same FSC recipe beyond the primary confirmation setting.
Source and complement ablations	Development comparisons, random and training-free sources, and confined complements	Attribute FSC gains to source construction, training, and directions outside complete entry support.

All in-distribution images in the geometry, diagnostic, and FSC studies come from the ImageNet validation set. The MAE readout is trained separately on the ImageNet training set (Appendix H), and the shifted datasets are described in Appendix D.1. The geometry study uses eight fit and two evaluation images per class; paired-view analyses use one evaluation anchor and two same-class controls per class. The diagnostic set is a separate class-balanced sample. Source learning uses 6,000 calibration and 2,000 tuning images from the geometry fit split, plus a 2,000-image development set. The matched-source analysis of Appendix F reuses this development set, and the confirmation study uses another 5,000 images (Appendix E). Later evaluation sets are disjoint from these earlier samples.

B.3 Augmentations and depth-dependent view advantage

This subsection defines the two augmentation families and the statistic that summarizes how view consistency changes with depth. The three models of Figure 2(a) are the AugReg checkpoints (Steiner et al., 2022) `vit_tiny_patch16_224`, `vit_small_patch16_224`, and `vit_base_patch16_224`, as provided through `timm` (Wightman, 2019), each with the pretrained tag `augreg_in21k_ft_in1k`. The Base checkpoint is the AugReg ViT-B used in the rest of the paper. Preprocessing uses each model's pretrained normalization. Mild views use random resized crops with scale $[0.8, 1.0]$ and horizontal flips. Strong views use scale $[0.6, 1.0]$, horizontal flips, RandAugment with two operations and magnitude 9, and color jitter $(0.2, 0.2, 0.2, 0.1)$. Two independently seeded views are generated per augmentation family. SVD is applied to the uncentered patch-token feature matrix, and subspaces are compared through Equation (2), which does not depend on the choice of basis within a subspace.

For canonical anchor i and layer ℓ, let $d^{\mathrm{cls}}_{i,\ell}$ be the average distance to its two same-class foreign-image controls and $d^{\mathrm{aug}}_{i,\ell}$ the distance between its two views. The reported advantage is $a_{i,\ell} = d^{\mathrm{cls}}_{i,\ell} - d^{\mathrm{aug}}_{i,\ell}$, and its mean over anchors, a_ℓ, is the shaded gap in Figure 2(b,c). The depth contrast is

$$\Delta a_i = \tfrac{1}{4}\sum_{\ell=8}^{11} a_{i,\ell} - \tfrac{1}{4}\sum_{\ell=0}^{3} a_{i,\ell}. \tag{B.2}$$

The contrast measures how much the separation between same-image views and same-class images grows with depth. Because both distances enter, the view advantage can grow through a larger

same-class distance as well as through a smaller view distance. The three widths of Figure 2(a) come from a separate sample of 10,000 images without anchors. Its final-block distances average 5,000 same-class and 5,000 different-class image pairs and the view pairs of 1,000 images. Its depth contrast is the same difference formed from mean distances, over 1,000 same-class pairs and the 1,000 view pairs, and 2,000 bootstrap draws resample images while keeping the layers of each pair together. For Tiny, Small, and Base, respectively, the changes are 0.1746, 0.2029, and 0.2678 under mild views, and 0.0920, 0.1329, and 0.1928 under strong views. Figure 2(b,c) and the parameter-state control below use the anchored sample.

B.4 Matched view control: pretrained versus random weights

To see how much of this geometry is already present at initialization, we compare the pretrained AugReg ViT-B with three random initializations of the same architecture (seeds 0, 1, and 2), at rank 16, across all 12 block outputs and both augmentation families. A random state is the architecture constructed by `timm` (Wightman, 2019) (version 1.0.11) with `pretrained=False` after seeding the random generators. All four parameter states receive the same 1,400 FP32 tensors: 600 unaugmented images and 800 augmented views, covering 200 classes with one anchor and two same-class controls per class. Three distinct images per class are needed, so 400 images come from the earlier evaluation split and 200 from its fit split. No source or classifier is fitted in this control, and confirmation images are excluded.

For each anchor and layer, the random-seed average is formed before comparison with the pretrained state. All 200 anchor/classes are resampled together across layers using 10,000 paired bootstrap draws. The contrasts compare pretrained-minus-random view advantage at zero-based blocks 8 to 11, the late-four mean, and the late-four-minus-early-four change for each augmentation. We report pointwise 95% percentile intervals. Table 6 gives the four aggregate contrasts:

Table 6: Matched parameter-state view control. Entries are pretrained minus the within-anchor mean over three random initializations; the 200 anchor classes are the resampling units. Bold highlights the pretrained-minus-random differences.

View	Contrast	Difference ↑	Pointwise 95% interval
Mild	Late-four gap	$+\mathbf{0.2424}$	$[0.2258, 0.2595]$
Mild	Late-minus-early gap	$+\mathbf{0.2298}$	$[0.2179, 0.2417]$
Strong	Late-four gap	$+\mathbf{0.2185}$	$[0.1975, 0.2400]$
Strong	Late-minus-early gap	$+\mathbf{0.1886}$	$[0.1763, 0.2008]$

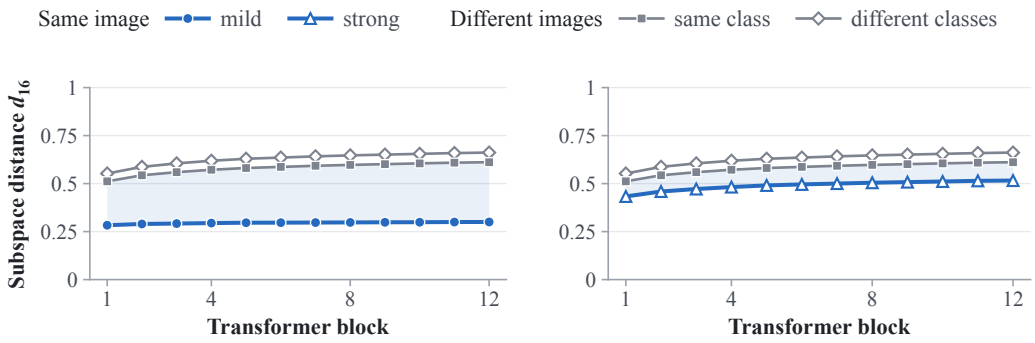
Figure 2(b,c) shows the pretrained distances and Figure 5 the random-initialization distances on the same scale. Same-image and same-class pairs share an anchor and enter the paired statistic; different-class curves use a fixed cyclic donor and serve as contextual references. Random initialization already produces view consistency, and in the first two blocks its mild-view advantage is even larger than that of the pretrained model, by 0.07 and 0.02. Pretraining mainly changes the late layers: in blocks 8 to 11 the pretrained checkpoint has a larger view advantage than the random states, and its advantage grows faster with depth (Table 6).

Numerical precision. SVDs use FP32 with an orthogonality tolerance of 5×10^{-4}. Bases that exceeded it were recomputed with a CPU FP64 SVD, which was needed for one random-state batch of eight inputs.

C Source construction and training

C.1 Geometric and predicted sources

Completion directions can come from three kinds of source: the input's own future, geometry shared across images, or a prediction of the future from the current entry. The *natural future* source of an image consists of the top 64 eigenvectors of the unnormalized Gram sum of its natural states at the entry and at both subsequent block outputs. Because it reads the image's own future, it serves only as



(a) Mild views (b) Strong views

Figure 5: **Random initializations also produce view-consistent subspaces.** Rank-16 distances under **(a)** mild and **(b)** strong views, averaged over the three random initializations for each of the 200 anchors in Figure 2(b,c). Axes, series encodings, and view-advantage shading match the main figure.

the privileged diagnostic of Section 5.2. A shared mean-projector source, G_{64}, instead averages the fit images' rank-64 future projectors and takes the top eigenspace of the average. The ridge comparator standardizes $z = \mathrm{mean}_{\mathrm{tokens}}(H^*)$ using fit statistics, includes a regularized intercept, and predicts the vectorized future projector. The matrix is symmetrized before taking its top eigenspace. Eight regularization values from 10^{-2} to 10^{5} are tried on a fixed tuning subset, and both models select 10^{3}.

The comparator recipes differ in when they select their eight candidates. The 64-column recipe first removes the full entry support $U = \mathrm{orth}([V_{64}, \mathbf{1}])$, then uses SVD to select eight directions. The FSC recipe starts with an eight-column source: reduced QR of A, two passes of subtraction of the U component, and reduced QR of the remainder produce the complement. Clean G_8 selects the shared bank's first eight columns *before* this same operation. Both recipes yield eight external directions, but they select from different candidate sets.

C.2 FSC OPTIMIZATION AND MATCHED INITIALIZATION

FSC trains one shared source per model with the recipe of Table 7, and evaluating the same source before training isolates what training adds.

Table 7: FSC training protocol for the two ViT-B models. Only the source A is optimized, and checkpoint selection uses teacher KL on the fixed tuning subset. All configuration extensions use the same optimization recipe.

Setting	Value
Source parameter	$A \in \mathbb{R}^{768\times 8}$, 6,144 scalars
Backbone / head	Frozen, with the original 1,000-class readout
Initialization	QR factor of (mean-projector top 8 + Gaussian noise), Equation (E.1)
Noise standard deviation	$0.01/\sqrt{768}$, with source seeds 0, 1, 2
Objective / temperature	$D_{\mathrm{KL}}(p_{\mathrm{natural}} \| p_A)$ at temperature 1
Optimizer	Adam, learning rate 0.003, weight decay 0
Budget	10 epochs per seed
Batches	Microbatch 8, effective batch 32
Selection	Minimum teacher KL on fixed 256-image tuning subset
Selected epochs, AugReg	8 / 10 / 9
Selected epochs, CLIP-FT	7 / 10 / 10
Matched comparison	Initialization (epoch 0) versus the selected source

Gradients propagate through the constrained suffix to A, including the effect of earlier constrained updates on later branches. Backbone weights, readout weights, common entries, entry bases, and

teacher logits remain fixed. Teacher KL drives optimization and selection. All eight-column sources use the same exclusion and rank rules. Inference, applied bases, and saved logits use FP32, while Gram and eigendecomposition steps use FP64 where specified.

The matched comparison on the development set evaluates, for each model, its three saved epoch-zero sources, the three sources selected from them, and clean G_8, without retraining or reselection. With source construction held fixed, the difference is the effect of optimization and selection (Appendix F).

C.3 Trainable source budget versus executed model size

Only the shared $D \times 8$ source is learned. For ViT-B this is 6,144 scalars, while the AugReg and CLIP-FT classification checkpoints from `timm` (Wightman, 2019), including their original 1,000-way heads, each contain approximately 86.6M parameters. Hence the trainable fraction is $100 \times 6{,}144/(86.6 \times 10^6) \approx 0.0071\%$, rounded to about 0.007% in the abstract. For ViT-S, the source contains 3,072 scalars. One source serves every input and every residual site of its suffix, and no matrix is learned per image or per layer.

The small trainable budget makes the direction-selection result concrete: the gains arise without updating the original model. Execution keeps the full-width attention and MLP branches and adds basis construction and update projection, so the parameter fraction measures what is learned. Training is also light: on one RTX 4090, a ViT-B source over two blocks trains for ten epochs in about four minutes, after a one-time pass of about 12 minutes that caches the entries, entry bases, and teacher logits of the 8,000 fit and tuning images.

D Additional controlled-intervention results

Contrast A1 isolates the cost of repeatedly restricting updates after an almost lossless entry projection. Contrast A2 compares two complete projection policies, and the entry-only-future arm separates what the future basis costs at the entry from what it costs during continuation. Common-entry comparisons then hold the starting tensor and budget fixed to test the directions themselves. Tables 9 and 10 give all diagnostic contrasts, and Table 8 gives the accuracies quoted in the main text. The raw controls in Table 11 test whether the conclusions depend on norm restoration, and Table 12 reports all four endpoints on the development set.

Table 8: Diagnostic-control accuracy (%) on 2,000 images. Complete-entry arms use 197 directions; common-entry arms use budget 64. T/R: token-norm/raw. Random is mean ± sample SD over five seeds. Bold highlights the persistent-Future arms.

	Complete entry ↑					Common entry ↑			
Setting	Natural	Entry (T)	Entry (R)	Fut. entry	Future	Free	Entry64	Random	Future
AugReg-B	85.05	69.35	67.80	84.90	**84.30**	84.25	52.80	54.73 ± 0.62	**79.55**
AugReg2-B	85.35	81.65	81.45	85.25	**85.05**	85.40	75.30	75.80 ± 0.48	**83.15**
CLIP-FT-B	86.15	53.00	71.25	85.75	**85.65**	84.80	3.90	4.86 ± 0.22	**82.50**
AugReg-S	82.00	73.65	73.95	81.35	**81.05**	80.55	57.15	57.41 ± 0.27	**74.50**
AugReg-L	86.45	85.50	85.30	86.55	**86.50**	85.45	84.70	84.64 ± 0.07	**85.90**
DINO-B / linear	79.25	71.80	74.20	78.45	**78.50**	72.30	63.55	65.59 ± 0.21	**70.40**
MAE-B / linear	67.05	0.50	0.70	54.20	**41.70**	58.95	0.30	0.29 ± 0.14	**1.80**
CLIP-B / zero-shot	68.40	23.55	37.50	68.05	**66.45**	68.50	11.40	11.31 ± 0.38	**54.30**

The complete-entry and common-entry diagnostics each use simultaneous intervals over their 16 CE comparisons. The development-set predictor comparison has six CE contrasts: predicted minus Entry64, predicted minus G_{64}, and predicted minus the three-seed learned-source mean, for each of two models. All six intervals exclude zero: the predictor improves on the first two comparators but remains below the learned source. Clean G_8, evaluated on the same development set, gives the construction-matched comparison in Appendix F.

Table 9: Complete-entry diagnostic. A1 compares persistent-entry with entry-only; A2 compares persistent-entry with persistent-future. All arms use token-norm, 197-column bases, and the final two blocks; CE effects are in nats with simultaneous 95% intervals. Bold highlights the CE effects.

Setting	Contrast	CE increase ↓	Simultaneous 95% interval
AugReg-B	A1	$\mathbf{+1.299}$	$[+1.220, +1.378]$
AugReg-B	A2	$\mathbf{+1.257}$	$[+1.180, +1.333]$
AugReg2-B	A1	$\mathbf{+0.429}$	$[+0.387, +0.472]$
AugReg2-B	A2	$\mathbf{+0.404}$	$[+0.362, +0.445]$
CLIP-FT-B	A1	$\mathbf{+4.541}$	$[+4.462, +4.620]$
CLIP-FT-B	A2	$\mathbf{+4.467}$	$[+4.391, +4.543]$
AugReg-S	A1	$\mathbf{+0.448}$	$[+0.397, +0.498]$
AugReg-S	A2	$\mathbf{+0.422}$	$[+0.372, +0.472]$
AugReg-L	A1	$\mathbf{+0.064}$	$[+0.047, +0.082]$
AugReg-L	A2	$\mathbf{+0.062}$	$[+0.045, +0.079]$
DINO-B / linear	A1	$\mathbf{+0.216}$	$[+0.170, +0.262]$
DINO-B / linear	A2	$\mathbf{+0.201}$	$[+0.157, +0.246]$
MAE-B / linear	A1	$\mathbf{+38.864}$	$[+37.660, +40.068]$
MAE-B / linear	A2	$\mathbf{+37.004}$	$[+35.822, +38.186]$
CLIP-B / zero-shot	A1	$\mathbf{+2.717}$	$[+2.588, +2.846]$
CLIP-B / zero-shot	A2	$\mathbf{+2.574}$	$[+2.442, +2.705]$

Table 10: Common-entry direction diagnostic. S compares privileged future with the random-complement mean, and B compares privileged future with Entry64. All arms use token-norm at nominal budget 64; CE effects are in nats with simultaneous 95% intervals. Bold highlights the CE effects.

Setting	Contrast	CE change ↓	Simultaneous 95% interval
AugReg-B	S	$\mathbf{-2.038}$	$[-2.131, -1.944]$
AugReg-B	B	$\mathbf{-2.149}$	$[-2.247, -2.052]$
AugReg2-B	S	$\mathbf{-0.658}$	$[-0.717, -0.598]$
AugReg2-B	B	$\mathbf{-0.670}$	$[-0.731, -0.609]$
CLIP-FT-B	S	$\mathbf{-5.148}$	$[-5.263, -5.033]$
CLIP-FT-B	B	$\mathbf{-5.253}$	$[-5.369, -5.138]$
AugReg-S	S	$\mathbf{-1.030}$	$[-1.115, -0.945]$
AugReg-S	B	$\mathbf{-1.065}$	$[-1.153, -0.977]$
AugReg-L	S	$\mathbf{-0.092}$	$[-0.115, -0.070]$
AugReg-L	B	$\mathbf{-0.093}$	$[-0.116, -0.071]$
DINO-B / linear	S	$\mathbf{-0.256}$	$[-0.312, -0.200]$
DINO-B / linear	B	$\mathbf{-0.372}$	$[-0.435, -0.308]$
MAE-B / linear	S	$\mathbf{-24.509}$	$[-25.826, -23.192]$
MAE-B / linear	B	$\mathbf{-23.447}$	$[-24.819, -22.075]$
CLIP-B / zero-shot	S	$\mathbf{-3.230}$	$[-3.426, -3.034]$
CLIP-B / zero-shot	B	$\mathbf{-3.281}$	$[-3.482, -3.081]$

The matched-source comparison uses its own two-contrast CE family with synchronized class-cluster bootstrap draws. CE receives intervals, while accuracy, KL, construction differences, and per-seed effects are reported as point values.

D.1 Distribution shift and window depth

To test whether the gap depends on in-distribution inputs or on the two-block window, we repeat the complete-entry control of Section 5 on four shifted datasets in all eight settings: ImageNetV2 matched-frequency (2,000 images, 1,000 classes), ImageNet-Sketch (1,995 images, 998 classes), ImageNet-R (2,000 images, 200 classes), and ImageNet-A (1,934 images, 200 classes) (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021a;b). For ImageNet-Sketch, perceptual-hash groups are kept together, groups with conflicting labels are dropped, and classes left with too few images

Table 11: Raw-projection controls. The complete-entry A1 contrast is evaluated in all eight settings, and common-entry S/B contrasts in three settings. CE effects are in nats with pointwise 95% intervals. Bold highlights the CE effects.

Setting	Contrast	CE change ↓	Pointwise 95% interval
AugReg-B	A1	$+\mathbf{1.256}$	$[+1.202, +1.311]$
AugReg2-B	A1	$+\mathbf{0.327}$	$[+0.299, +0.356]$
CLIP-FT-B	A1	$+\mathbf{3.665}$	$[+3.610, +3.722]$
AugReg-S	A1	$+\mathbf{0.394}$	$[+0.359, +0.430]$
AugReg-L	A1	$+\mathbf{0.077}$	$[+0.063, +0.091]$
DINO-B / linear	A1	$+\mathbf{0.160}$	$[+0.133, +0.187]$
MAE-B / linear	A1	$+\mathbf{32.676}$	$[+31.939, +33.430]$
CLIP-B / zero-shot	A1	$+\mathbf{1.587}$	$[+1.509, +1.666]$
AugReg-B	S	$-\mathbf{1.781}$	$[-1.843, -1.719]$
AugReg-B	B	$-\mathbf{1.879}$	$[-1.945, -1.815]$
MAE-B / linear	S	$-\mathbf{14.868}$	$[-15.570, -14.169]$
MAE-B / linear	B	$-\mathbf{13.918}$	$[-14.640, -13.200]$
CLIP-B / zero-shot	S	$-\mathbf{1.690}$	$[-1.796, -1.581]$
CLIP-B / zero-shot	B	$-\mathbf{1.721}$	$[-1.827, -1.610]$

Table 12: Development-set endpoints. A_0 and FSC are mean ± sample SD over three seeds. CE and teacher KL are in nats; accuracy and flips (images whose predicted class differs from the natural model's) are percentages. Bold highlights FSC.

Model	Policy	CE ↓	KL ↓	Acc. ↑	Flips ↓
AugReg-B	Natural model	0.555	0.000	84.05	0.00
	Free from H^*	0.576	0.029	83.55	4.40
	Entry64	2.886	2.311	52.85	43.85
	Shared geometric G_{64}	2.925	2.344	54.00	42.50
	Clean geometric G_8	3.408	2.814	36.45	61.50
	Predicted future	2.194	1.628	58.50	36.70
	Matched initialization A_0	3.4084 ± 0.0002	2.8147 ± 0.0001	36.30 ± 0.09	61.62 ± 0.03
	FSC (selected A_*)	**1.463** ± 0.003	**0.906** ± 0.003	**70.28** ± 0.37	**23.62** ± 0.42
	Privileged future	0.797	0.245	80.35	12.15
CLIP-FT-B	Natural model	0.626	0.000	84.55	0.00
	Free from H^*	0.674	0.028	84.15	2.50
	Entry64	6.826	5.661	3.65	96.35
	Shared geometric G_{64}	5.862	4.765	22.35	76.95
	Clean geometric G_8	2.633	1.788	71.85	22.65
	Predicted future	4.841	3.828	47.35	49.85
	Matched initialization A_0	2.634 ± 0.009	1.788 ± 0.008	71.88 ± 0.03	22.65 ± 0.09
	FSC (selected A_*)	**1.509** ± 0.001	**0.812** ± 0.001	**76.62** ± 0.19	**16.50** ± 0.09
	Privileged future	1.566	0.819	82.55	8.20

are not padded. For AugReg-B, MAE-B, and zero-shot CLIP-B, we also evaluate ImageNet-C (Hendrycks & Dietterich, 2019) on the diagnostic images under all 15 corruption types and five severities, averaging the 75 conditions of each image before aggregating over images.

The reference for each shifted input is the unmodified model on that same input, and the primary metric is KL divergence over the full 1,000-way output. This avoids mixing the effect of distribution shift with the effect of the constraint. The natural-shift family has 64 A1/A2 contrasts, and ImageNet-C adds six more; all intervals exclude zero (Table 13). The persistent-future arm stays within 4 accuracy points of the natural model everywhere except MAE-B, where it recovers about half of the loss, matching the in-distribution pattern.

Fidelity and task performance can disagree on shifted data, and the label space matters. For AugReg-L on ImageNet-R, the persistent-entry constraint moves the predictions away from the natural ones

(KL 0.1709 nats) while CE *falls* by 0.1605 nats and accuracy rises by 0.7 points. Restricted to the dataset's 200-class label subset, CE instead rises by 0.0322 nats. The two halves of the data explain the sign: on the 1,014 images that the natural model classifies correctly CE rises by 0.178 nats, and on the 986 that it misclassifies CE falls by 0.508 nats. Perturbing a wrong prediction can lower CE without preserving anything, so KL to the natural output is the primary metric under shift.

Table 13: Complete-entry control under distribution shift. Natural is unmodified-model accuracy; Loss E/F is the accuracy loss of persistent-entry or persistent-future. KL is divergence from the same image's natural prediction, with simultaneous 95% intervals within the natural-shift and ImageNet-C families. Bold highlights the paired KL excess of Entry over Future.

Dataset	Natural ↑	Loss E ↓	Loss F ↓	KL E ↓ (95% CI)	KL E − F ↓ (95% CI)
AugReg-B					
ImageNetV2	74.4	16.3	0.9	1.396 [1.327, 1.466]	**1.350** [1.281, 1.419]
ImageNet-Sketch	43.9	15.6	0.8	1.476 [1.411, 1.540]	**1.416** [1.351, 1.480]
ImageNet-R	43.3	18.6	0.9	1.636 [1.559, 1.713]	**1.557** [1.480, 1.634]
ImageNet-A	24.3	15.5	1.6	1.743 [1.661, 1.826]	**1.643** [1.562, 1.724]
ImageNet-C	67.0	19.3	0.7	1.536 [1.500, 1.571]	**1.482** [1.447, 1.518]
AugReg2-B					
ImageNetV2	74.9	4.8	0.1	0.402 [0.374, 0.430]	**0.389** [0.362, 0.417]
ImageNet-Sketch	46.9	6.4	0.5	0.559 [0.529, 0.590]	**0.543** [0.513, 0.573]
ImageNet-R	44.6	7.2	−0.3	0.613 [0.574, 0.653]	**0.593** [0.554, 0.632]
ImageNet-A	27.0	9.2	1.3	0.757 [0.715, 0.799]	**0.728** [0.688, 0.769]
CLIP-FT-B					
ImageNetV2	75.2	32.0	0.2	4.025 [3.954, 4.096]	**3.987** [3.915, 4.058]
ImageNet-Sketch	52.4	17.7	−0.1	3.271 [3.190, 3.352]	**3.212** [3.130, 3.294]
ImageNet-R	50.2	21.1	0.3	3.274 [3.165, 3.383]	**3.215** [3.104, 3.326]
ImageNet-A	27.3	14.4	0.5	3.219 [3.130, 3.308]	**3.148** [3.058, 3.237]
AugReg-S					
ImageNetV2	70.1	8.6	0.6	0.525 [0.487, 0.563]	**0.497** [0.460, 0.534]
ImageNet-Sketch	33.8	6.8	0.7	0.668 [0.636, 0.700]	**0.636** [0.605, 0.668]
ImageNet-R	32.5	8.1	−0.2	0.738 [0.699, 0.777]	**0.691** [0.653, 0.729]
ImageNet-A	12.8	4.2	−0.2	0.838 [0.798, 0.879]	**0.776** [0.737, 0.814]
AugReg-L					
ImageNetV2	75.7	0.3	−0.5	0.076 [0.068, 0.083]	**0.072** [0.065, 0.079]
ImageNet-Sketch	54.5	0.6	−0.1	0.130 [0.121, 0.140]	**0.128** [0.118, 0.137]
ImageNet-R	50.7	−0.7	−0.7	0.171 [0.157, 0.185]	**0.165** [0.152, 0.179]
ImageNet-A	32.6	−1.5	−0.6	0.164 [0.153, 0.175]	**0.154** [0.144, 0.165]
DINO-B / linear					
ImageNetV2	66.8	4.8	0.6	0.315 [0.288, 0.342]	**0.296** [0.270, 0.322]
ImageNet-Sketch	27.3	2.3	−0.5	0.450 [0.430, 0.471]	**0.432** [0.412, 0.452]
ImageNet-R	27.3	2.5	−0.5	0.468 [0.445, 0.492]	**0.441** [0.418, 0.464]
ImageNet-A	5.6	0.4	0.4	0.609 [0.576, 0.643]	**0.568** [0.536, 0.600]
MAE-B / linear					
ImageNetV2	53.9	53.1	23.1	39.36 [38.34, 40.37]	**37.18** [36.18, 38.18]
ImageNet-Sketch	15.9	15.6	6.2	41.47 [40.76, 42.18]	**39.25** [38.54, 39.97]
ImageNet-R	17.4	17.2	9.0	40.84 [39.98, 41.71]	**38.05** [37.19, 38.91]
ImageNet-A	1.5	1.4	0.3	40.70 [39.83, 41.57]	**37.54** [36.68, 38.40]
ImageNet-C	30.1	29.8	13.6	43.32 [42.98, 43.67]	**40.95** [40.60, 41.30]
CLIP-B / zero-shot					
ImageNetV2	60.9	39.6	2.6	2.670 [2.560, 2.781]	**2.519** [2.408, 2.630]
ImageNet-Sketch	50.3	33.8	2.3	2.563 [2.464, 2.662]	**2.415** [2.316, 2.514]
ImageNet-R	57.3	36.6	3.3	2.430 [2.308, 2.552]	**2.236** [2.114, 2.357]
ImageNet-A	29.4	20.2	3.9	2.555 [2.433, 2.676]	**2.316** [2.194, 2.438]
ImageNet-C	44.7	30.7	3.0	2.368 [2.321, 2.414]	**2.189** [2.142, 2.236]

Direction choice was also tested under shift, in the same three settings and on the four natural-shift datasets (Table 14). The ordering of Section 5.2 is unchanged: future-derived directions lower CE relative to both Entry64 and random complements in all 24 comparisons, whose pointwise 95% intervals exclude zero, and random complements again stay close to Entry64. On ImageNetV2,

AugReg-B rises from 42.0% to 70.1% against 73.4% for free continuation. In MAE-B, CE again improves by more than 20 nats while accuracy stays below 2%.

Table 14: Direction choice under distribution shift. For each shifted dataset and tested setting, common-entry arms share H^*, 64 token-norm columns, the last-two-block suffix, and the full 1,000-way output. Random accuracy is mean $\pm$ sample SD over five seeds. The last two columns report the CE change of Future against Entry64 and Random, in nats, with pointwise 95% intervals. Bold highlights Future accuracy.

	Accuracy (%) ↑				CE change of Future ↓ (95% CI)	
Dataset	Free ↑	Entry64 ↑	Random ↑	Future ↑	vs. Entry64	vs. Random
AugReg-B						
ImageNetV2	73.4	42.0	43.7 ± 0.4	**70.1**	-2.09 $[-2.16, -2.02]$	-1.98 $[-2.05, -1.91]$
ImageNet-Sketch	44.4	19.2	18.9 ± 0.2	**41.3**	-1.85 $[-1.93, -1.77]$	-1.79 $[-1.87, -1.71]$
ImageNet-R	43.2	14.5	14.8 ± 0.1	**40.3**	-1.82 $[-1.95, -1.69]$	-1.75 $[-1.88, -1.62]$
ImageNet-A	23.3	4.9	5.3 ± 0.2	**18.1**	-1.29 $[-1.42, -1.16]$	-1.24 $[-1.36, -1.11]$
MAE-B / linear						
ImageNetV2	45.8	0.3	0.3 ± 0.1	**1.6**	-22.9 $[-23.9, -22.0]$	-23.6 $[-24.5, -22.7]$
ImageNet-Sketch	13.6	0.2	0.2 ± 0.1	**0.4**	-21.1 $[-22.1, -20.2]$	-21.5 $[-22.4, -20.6]$
ImageNet-R	13.7	0.3	0.2 ± 0.1	**0.3**	-20.1 $[-21.7, -18.6]$	-20.8 $[-22.3, -19.4]$
ImageNet-A	1.6	0.2	0.2 ± 0.1	**0.3**	-20.3 $[-22.0, -18.6]$	-21.1 $[-22.7, -19.4]$
CLIP-B / zero-shot						
ImageNetV2	60.7	10.3	10.4 ± 0.3	**46.6**	-3.09 $[-3.23, -2.94]$	-3.01 $[-3.15, -2.86]$
ImageNet-Sketch	49.7	7.7	8.5 ± 0.2	**39.3**	-2.99 $[-3.13, -2.85]$	-2.94 $[-3.08, -2.80]$
ImageNet-R	57.1	12.7	13.2 ± 0.7	**45.1**	-2.35 $[-2.58, -2.12]$	-2.33 $[-2.56, -2.10]$
ImageNet-A	29.8	4.5	5.0 ± 0.3	**14.5**	-1.38 $[-1.65, -1.10]$	-1.35 $[-1.62, -1.07]$

A development analysis varies the number of replaced directions, $k \in \{1, 4, 8, 16\}$, in the same three settings on 128 images from the evaluation split of the geometry study (Table 15). A single future-derived direction already lowers CE relative to the random complement, and the advantage grows with k: for AugReg-B the gap to free continuation falls from 1.64 nats at $k = 1$ to 0.30 at $k = 8$ and 0.15 at $k = 16$, while random complements stay near 2.4 nats. Most of the benefit is reached by $k = 8$, the value used throughout the paper. Because the shared entry changes with k, comparisons are within rows.

Table 15: Number of replaced directions in the common-entry control. At nominal budget 64, k entry directions are replaced on 128 images from the evaluation split of the geometry study. Gap to free is arm CE minus same-entry free CE; the last column compares Future with the random-complement mean, with pointwise 95% intervals. Bold highlights the Future gap.

		Gap to free		
Setting	k	Random gap ↓	Future gap ↓	Future − Random ↓ (95% CI)
AugReg-B	1	2.35	**1.64**	-0.71 $[-0.90, -0.53]$
	4	2.35	**0.67**	-1.68 $[-1.89, -1.47]$
	8	2.36	**0.30**	-2.06 $[-2.27, -1.85]$
	16	2.45	**0.15**	-2.30 $[-2.53, -2.07]$
MAE-B / linear	1	47.0	**41.8**	-5.2 $[-7.3, -3.1]$
	4	46.7	**26.8**	-19.9 $[-22.8, -17.0]$
	8	46.4	**21.9**	-24.5 $[-27.6, -21.3]$
	16	47.6	**21.0**	-26.6 $[-29.8, -23.4]$
CLIP-B / zero-shot	1	3.73	**3.20**	-0.53 $[-0.81, -0.26]$
	4	3.74	**1.93**	-1.81 $[-2.27, -1.36]$
	8	3.69	**0.91**	-2.79 $[-3.24, -2.35]$
	16	3.92	**0.65**	-3.27 $[-3.74, -2.82]$

The four-block condition shows how the gap changes with the number of constrained blocks. For AugReg-L, whose two-block loss is the smallest of the eight settings, this condition constrains the last

four blocks: the entry is the output of block 19, and all eight updates of blocks 20 to 23 are restricted (zero-based indices). Table 16 shows that the accuracy loss of the persistent-entry arm grows from 0.95 to 4.10 points and its CE increase from 0.064 to 0.302 nats, while the persistent-future arm loses 0.30 points. FSC over four blocks is evaluated in Appendix E.

Table 16: Window-depth control in AugReg-L. Complete-entry controls use 197-column bases and token-norm on the 2,000 diagnostic images. E and F denote the persistent-entry and persistent-future arms. Loss is accuracy lost relative to the natural model; CE reports A1 and A2 in nats, with the simultaneous intervals of Table 9 for two blocks and pointwise intervals for four. Bold highlights Future loss.

Blocks	Updates	Loss E↓	Loss F↓	CE A1↓ (95% CI)	CE A2↓ (95% CI)
2	4	0.95	$\mathbf{-0.05}$	$+0.064$ $[+0.047, +0.082]$	$+0.062$ $[+0.045, +0.079]$
4	8	4.10	**0.30**	$+0.302$ $[+0.274, +0.330]$	$+0.295$ $[+0.268, +0.323]$

The gap also appears away from the end of the network. A mid-depth condition moves the two-block window to blocks 5 and 6 of the 12-block models (blocks 11 and 12 of AugReg-L; zero-based indices), with the entry at the output of block 4 (block 10) and the remaining blocks continuing freely to the readout, on the same 2,000 diagnostic images with the same three arms and 197-column bases (Table 17). All sixteen A1 and A2 intervals, simultaneous within the family, exclude zero. In six of the eight settings the persistent-entry arm loses more accuracy than at the last two blocks, by a wide margin in AugReg-B (41.4 against 15.7 points), AugReg-S (29.8 against 8.4), and zero-shot CLIP-B (68.1 against 44.9), whereas CLIP-FT-B and MAE-B lose somewhat less. The persistent-future basis again keeps the loss within 1.4 points except in MAE-B (7.4). The five free blocks after the window (eleven in AugReg-L) do not repair what the entry space lacked.

Table 17: Complete-entry control at mid depth. The entry is the output of block 4 and blocks 5–6 are constrained (zero-based; blocks 11–12 in AugReg-L), after which the remaining blocks continue freely. E and F denote the persistent-entry and persistent-future arms. Loss is accuracy lost in points, with the E loss of the last-two-block window for reference, and CE effects are in nats with simultaneous 95% intervals. Bold highlights mid-depth Future loss.

		Accuracy loss↓			CE effect, blocks 5–6↓ (95% CI)	
Setting	Natural↑	E, 5–6	F, 5–6	E, last two	A1	A2
AugReg-B	85.1	41.4	**0.7**	15.7	$+2.25$ $[+2.10, +2.41]$	$+2.24$ $[+2.08, +2.39]$
AugReg2-B	85.4	9.3	$\mathbf{-0.1}$	3.7	$+0.46$ $[+0.39, +0.53]$	$+0.45$ $[+0.39, +0.52]$
CLIP-FT-B	86.2	23.1	**0.5**	33.2	$+1.14$ $[+1.03, +1.25]$	$+1.12$ $[+1.00, +1.23]$
AugReg-S	82.0	29.8	**0.6**	8.4	$+1.52$ $[+1.39, +1.65]$	$+1.48$ $[+1.35, +1.61]$
AugReg-L	86.5	1.4	**0.4**	1.0	$+0.04$ $[+0.02, +0.06]$	$+0.04$ $[+0.02, +0.06]$
DINO-B / linear	79.3	8.2	**1.4**	7.5	$+0.31$ $[+0.27, +0.36]$	$+0.30$ $[+0.25, +0.35]$
MAE-B / linear	67.1	62.7	**7.4**	66.6	$+7.24$ $[+6.93, +7.56]$	$+6.85$ $[+6.54, +7.16]$
CLIP-B / zero-shot	68.4	68.1	**1.3**	44.9	$+7.71$ $[+7.48, +7.94]$	$+7.60$ $[+7.37, +7.83]$

D.2 Discarded update energy, the two projection rules, and the MAE-B readout

How much of an update the entry basis discards. For each of the four natural updates of the two-block window, we measure the share of its squared Frobenius norm that lies outside the support of the complete entry basis, $\mathrm{span}(V_{197}, \mathbf{1})$. Averaged over the 2,000 diagnostic images, this share ranges from 17% to 75% across the eight settings and four sites (46%, 66%, 25%, and 43% in AugReg-B). A basis that holds the entire entry thus misses between a sixth and three quarters of what the next two blocks write, the situation that Equation (4) describes. The future basis also discards energy, between 3% and 43% by the same measure, yet it supports the computation (Table 8). In zero-shot CLIP-B it discards 43% of the first attention update where the entry basis discards 64%, and it loses 2.0 accuracy points where the entry basis loses 44.9. Discarded energy therefore does not predict the damage, in line with Proposition 2.

Token-norm and raw projection. Token-norm rescales the retained part of each update to the original centered norm. Along the constrained trajectory, the aggregate factor $1/\sqrt{1-\text{share}}$ has site means between 1.1 and 2.0 across the eight settings and is at most 2.4 for a single image. Measured per token, the factor is below 3 in all eight settings (99.9th percentile at most 2.64). Over every token-norm arm of the diagnostics and of the main FSC comparison, the smallest projected centered norm is 0.54, so the clamp of Equation (B.1) never acts. The two rules give accuracy losses within 2.4 points of each other in six settings. In the two CLIP settings raw projection loses less, 14.9 points against 33.2 in CLIP-FT-B and 30.9 against 44.9 in zero-shot CLIP-B (Table 8). CLIP-FT-B is rescaled no more than AugReg2-B, where the two rules agree (site means of 1.2 to 1.7 against 1.4 to 1.7), so the difference reflects how strongly a model responds to rescaled updates and not the size of the rescaling alone. Under either rule, the interval of the CE increase excludes zero in all eight settings (Tables 9 and 11, the latter with pointwise intervals).

Why MAE-B is extreme. MAE-B loses 66.6 points and its CE rises by 38.9 nats, an order of magnitude more than in any other setting, and the reason lies in its readout. The linear probe of the official MAE recipe standardizes every channel of the final class token with frozen BatchNorm statistics before the linear layer (Appendix H), and the standard deviations it divides by are small. The class token has a median norm of 16.5 and the running mean a norm of 15.9, so most of the token is an offset shared by all images, and the median channel standard deviation is 0.063. Confining the four updates changes the class token by 0.70 of its norm (median over images), less than in AugReg-B (0.96) or AugReg2-B (1.38), but after standardization the change is 3.7 times the norm of the standardized token. MAE-B is accordingly the only setting whose logits grow under the constraint: the median largest absolute logit rises from 17 to 53, whereas it falls in the other seven. The future basis changes the class token by only 0.11 and still costs 25.4 points. MAE-B therefore shows the same gap through a readout that tolerates far less feature error, and its CE values mainly reflect inflated logits.

E Held-out confirmation and configuration extensions

E.1 Held-out images and fixed comparisons

The confirmation study tests whether the selected sources generalize beyond the images used during development. The confirmation set contains 5,000 ImageNet validation images, five per class. It excludes images from the fit, tuning, development, diagnostic, and view-control pools, together with exact file and decoded-pixel duplicates.

These sources were selected by teacher KL on the fixed tuning subset, before any confirmation evaluation. The main comparison evaluates the six checkpoints *selected during development*, three per ViT-B model, without retraining or reselection.

The extensions vary model width, nominal budget, suffix length, and projection rule under the same learning recipe. Seven standard configurations cover AugReg ViT-S, four ViT-B budget settings, and two ViT-B four-block settings; two raw-projection configurations are also included. Three sources per extension give 27 new training runs. Together with the six existing sources, the study contains 11 configurations and 33 checkpoint-initialization pairs, and Appendix E.5 adds 21 configurations and 63 pairs trained later with the same recipe. Each configuration evaluates natural, free-continuation, entry, clean-G_8, initialization, and trained-source arms on the same 5,000 images.

E.2 Source construction across budgets, widths, and windows

When the model, the budget, or the suffix changes, sources are built as follows. For each model and suffix window, fit images provide a geometric initialization bank. On each fit image, an unnormalized Gram sum uses the natural entry and natural block-boundary outputs in that window, and its top 64 directions define a projector. The leading eigendirections of the mean of these fit-image projectors form the shared bank. The two-block bank uses zero-based block outputs 9, 10, and 11, and the four-block bank uses outputs 7 through 11. At a fixed model and window, the same bank is reused across nominal budgets, and raw projection reuses the corresponding natural two-block bank. Thus changing the window changes the bank, whereas changing the nominal budget does not introduce another independently selected bank.

Let B_8 be this bank's first eight columns. For source seed s, the saved initialization is

$$A_0^{(s)} = \mathrm{QR}\left(B_8 + \frac{0.01}{\sqrt{D}} Z^{(s)}\right), \qquad Z_{ij}^{(s)} \sim \mathcal{N}(0,1), \tag{E.1}$$

where QR returns the orthonormal factor. Gaussian sampling and this QR are performed on CPU. Clean G_8 uses the unperturbed B_8 through the same eight-column construction. A source is first QR-orthogonalized, then residualized twice against $\mathrm{span}(V_r, \mathbf{1})$, rank-checked, and QR-orthogonalized again before concatenation with V_{r-8}. The mean direction remains excluded for both raw and token-norm variants. This selects candidates before residualization, in contrast to the 64-column-then-SVD recipe in Appendix C. Each source contains $D \times 8$ scalars: 6,144 for ViT-B and 3,072 for ViT-S.

The common entry is recomputed once per configuration and image as $H^* = T_{\rho, P_{r-8}}(H_{\mathrm{entry}})$. Every arm in that cell consumes the same cached tensor. The free arm leaves all subsequent updates unrestricted, the entry-r arm applies the original V_r, and completion arms apply $[V_{r-8}, C_A]$. For the two-block window, the four projected residual additions occur in blocks 10 and 11, using zero-based indices. The four-block window projects all eight attention/MLP additions in blocks 8 to 11. Updates are transformed before addition, as in Equation (1). The natural teacher is evaluated separately without an entry intervention.

Learning uses the original 6,000 fit images, frozen model and readout, ten Adam epochs, learning rate 0.003, zero weight decay, temperature one, microbatch eight, and effective batch 32. Natural-teacher KL is differentiated through the frozen suffix to the source alone. The 256-image tuning subset selects the minimum-KL checkpoint over epochs 0 to 10, breaking ties toward the earliest epoch. Source selection uses neither labels nor confirmation images. At evaluation, the selected source and the current entry basis determine the applied space, and the natural forward pass of a test image provides the fidelity reference.

E.3 Estimands and uncertainty

This subsection defines the aggregation, contrasts, and intervals. For image i in class c, let $m_{ci}^{a,s}$ be an outcome metric for arm a and source seed s. We form

$$\bar{m}_{ci}^{a} = \frac{1}{3}\sum_{s=0}^{2} m_{ci}^{a,s}, \qquad \bar{m}_{c}^{a} = \frac{1}{5}\sum_{i=1}^{5} \bar{m}_{ci}^{a}, \qquad \widehat{m}^{a} = \frac{1}{1000}\sum_{c=1}^{1000} \bar{m}_{c}^{a}. \tag{E.2}$$

Seed-independent references enter without a seed average. No logits or predictions are ensembled. With five images in every class, class-macro and image-micro metric means coincide. CE and KL are in nats, and accuracy and teacher agreement are percentages when displayed.

Seed variability versus sampling uncertainty. Tabulated mean $\pm$ SD uses the sample standard deviation of the three source-specific class-macro metrics, $s_{\mathrm{seed}} = \sqrt{\sum_{s=0}^{2}(m_s - \bar{m})^2/(3-1)}$. For a contrast, we first subtract the two metrics within each matched seed and then compute its mean and SD. The five-seed random diagnostic uses the same definition with denominator $5-1$. These SDs describe variation across the saved sources on fixed evaluation images. Statistical significance uses the class-bootstrap 95% confidence intervals defined below, conditional on those sources; bracketed ranges give their endpoints. Deterministic arms have no seed SD, and neither do ratios of seed-mean gains; the retained shares of Table 27 carry bootstrap intervals instead.

The two CE contrasts in each configuration are $\Delta_{\mathrm{train}} = \widehat{\mathrm{CE}}(A_0) - \widehat{\mathrm{CE}}(A_*)$ and $\Delta_{G_8} = \widehat{\mathrm{CE}}(G_8) - \widehat{\mathrm{CE}}(A_*)$. Positive values favor the selected source. The remaining deficit is $G_{\mathrm{free}} = \widehat{\mathrm{CE}}(A_*) - \widehat{\mathrm{CE}}(\mathrm{free})$. Together the three quantities separate what training adds to its initialization, what it adds beyond a geometric construction, and what is still lost relative to unrestricted continuation.

For uncertainty, we resample classes together across policies and source seeds, keeping each within-image comparison intact. The four frozen-checkpoint CE contrasts (two models by two comparators) use 10,000 common draws of the 1,000 classes with replacement. If $\widehat{\Delta}_j$ is a contrast and $\widehat{\sigma}_j$ the standard deviation of its bootstrap estimates, the common critical value is the 95th percentile of

$$\max_{j \in \{1,2,3,4\}} \frac{|\widehat{\Delta}_j^{(b)} - \widehat{\Delta}_j|}{\widehat{\sigma}_j}. \tag{E.3}$$

It equals 2.2243 in these draws, so the simultaneous intervals are $\widehat{\Delta}_j \pm 2.2243\widehat{\sigma}_j$. Standard and raw extensions use pointwise class-bootstrap intervals from the 2.5th to the 97.5th percentile. Figure 6 shows positive matched CE reductions in all 11 configurations.

Table 18: Confirmation CE reductions (nats; positive favors FSC). Each estimate is mean ± sample SD over three paired seeds. Separate columns give class-bootstrap 95% CIs: simultaneous for the first two rows, pointwise elsewhere. T: token-norm; R: raw. Bold highlights the matched training gain.

				$A_0 - A_*$ CE↑		$G_8 - A_*$ CE↑	
Model	r	Blocks	Rule	Mean ± SD	95% CI	Mean ± SD	95% CI
AugReg-B	64	2	T	**1.962** ± 0.004	1.903–2.020	1.961 ± 0.004	1.903–2.019
CLIP-FT-B	64	2	T	**1.116** ± 0.008	1.083–1.149	1.115 ± 0.002	1.082–1.148
AugReg-S	64	2	T	**0.456** ± 0.002	0.424–0.487	0.459 ± 0.003	0.427–0.490
AugReg-B	32	2	T	**2.356** ± 0.003	2.306–2.406	2.358 ± 0.003	2.308–2.409
AugReg-B	128	2	T	**1.595** ± 0.005	1.548–1.642	1.593 ± 0.006	1.546–1.639
CLIP-FT-B	32	2	T	**2.728** ± 0.016	2.678–2.779	2.725 ± 0.017	2.674–2.775
CLIP-FT-B	128	2	T	**0.707** ± 0.005	0.687–0.726	0.7078 ± 0.0005	0.688–0.728
AugReg-B	64	4	T	**2.353** ± 0.009	2.294–2.412	2.350 ± 0.007	2.291–2.409
CLIP-FT-B	64	4	T	**1.967** ± 0.010	1.926–2.008	1.963 ± 0.002	1.922–2.003
AugReg-B	64	2	R	**1.290** ± 0.011	1.248–1.332	1.291 ± 0.010	1.249–1.333
CLIP-FT-B	64	2	R	**0.503** ± 0.001	0.484–0.523	0.503 ± 0.003	0.483–0.522

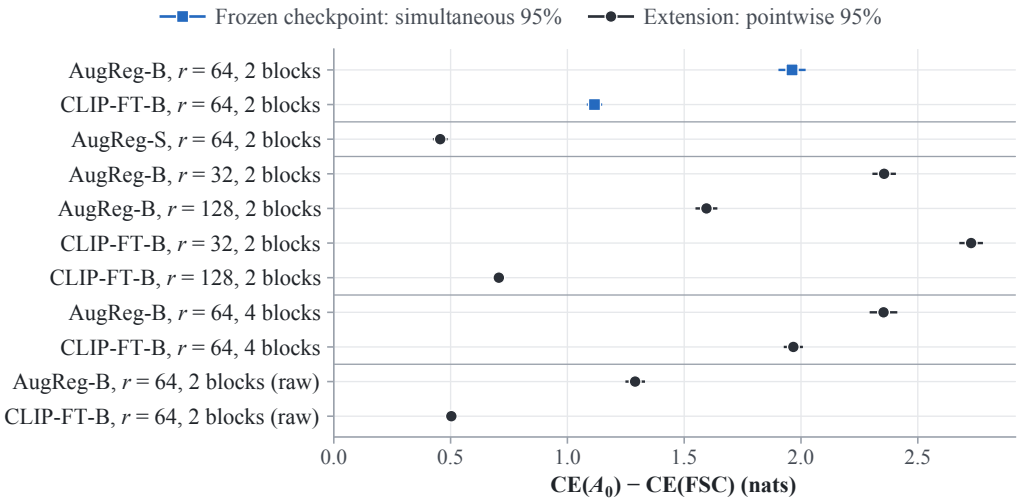


Figure 6: **FSC improves its matched initialization in every configuration.** Values are $\mathrm{CE}(A_0) - \mathrm{CE}(\mathrm{FSC})$ on the same 5,000 confirmation images, so positive values favor FSC. Squares are the two A_0 contrasts in the four-member simultaneous 95% family; circles are pointwise 95% intervals for extensions. Configurations are grouped by varied factor, and clean-G_8 contrasts appear in Table 18.

E.4 Absolute performance and remaining gaps

Table 19 gives the absolute accuracy of every arm in all eleven configurations, and Table 20 gives the fidelity of the selected sources together with their remaining CE gap to free continuation.

Figures 7 and 4 show two things at once: the improvement over the initialization and the distance that remains to free continuation. Every configuration has its own common entry and separately trained source, and the four-block suffix also has its own entry location and fitted bank, so each point is a complete configuration. In the four-block setting, gains of 27.71 and 23.13 accuracy points still leave

Table 19: Absolute confirmation accuracy (%). A_0 and FSC are mean ± sample SD over three seeds. Configurations give model, budget, suffix length, and rule (T: token-norm; R: raw). Each free reference uses the same entry as its constrained arms. Bold highlights FSC.

Configuration	Natural↑	Free↑	Entry-r↑	G_8↑	A_0↑	FSC↑
AugReg-B, 64, 2, T	83.98	83.00	50.64	34.38	34.34 ± 0.19	**70.53** ± 0.26
CLIP-FT-B, 64, 2, T	85.04	84.50	4.02	72.88	72.79 ± 0.19	**77.60** ± 0.24
AugReg-S, 64, 2, T	80.56	79.42	56.40	61.82	61.93 ± 0.22	**71.04** ± 0.11
AugReg-B, 32, 2, T	83.98	76.10	26.86	28.38	28.34 ± 0.08	**66.91** ± 0.17
AugReg-B, 128, 2, T	83.98	83.84	61.38	48.56	48.31 ± 0.19	**74.24** ± 0.07
CLIP-FT-B, 32, 2, T	85.04	81.24	2.98	43.90	44.04 ± 0.21	**66.95** ± 0.49
CLIP-FT-B, 128, 2, T	85.04	85.06	20.34	78.26	78.37 ± 0.15	**80.46** ± 0.03
AugReg-B, 64, 4, T	83.98	82.98	0.50	2.70	2.69 ± 0.06	**30.41** ± 0.51
CLIP-FT-B, 64, 4, T	85.04	85.02	0.78	14.08	13.91 ± 0.11	**37.04** ± 0.32
AugReg-B, 64, 2, R	83.98	83.08	52.94	48.58	48.73 ± 0.08	**68.91** ± 0.28
CLIP-FT-B, 64, 2, R	85.04	84.42	15.70	76.68	76.53 ± 0.13	**77.76** ± 0.16

Table 20: Fidelity and remaining deficits, mean ± sample SD over three source seeds. CE and teacher KL are in nats; agreement is in %. ΔKL is matched A_0 minus FSC, computed within seed before averaging. Bold highlights teacher KL, the training objective.

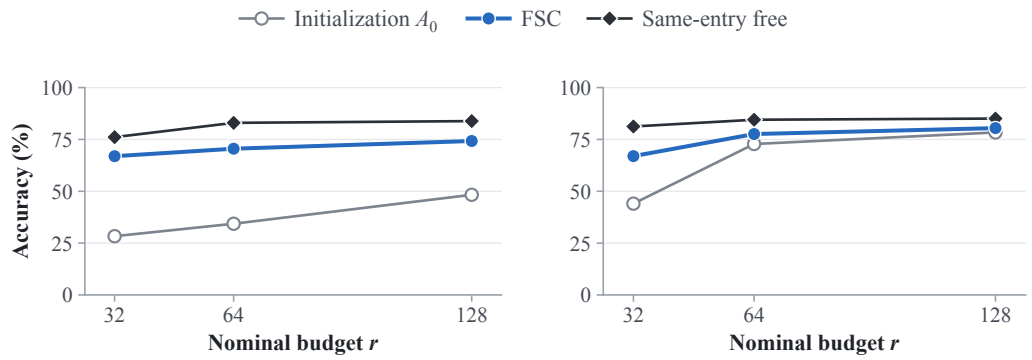
Configuration	CE↓	Teacher KL↓	Agreement↑	CE gap to free↓	ΔKL↑
AugReg-B, 64, 2, T	1.509 ± 0.004	**0.925** ± 0.001	76.12 ± 0.19	0.894 ± 0.004	1.941 ± 0.001
CLIP-FT-B, 64, 2, T	1.500 ± 0.002	**0.807** ± 0.002	83.76 ± 0.09	0.818 ± 0.002	0.967 ± 0.007
AugReg-S, 64, 2, T	1.229 ± 0.003	**0.565** ± 0.003	77.73 ± 0.08	0.468 ± 0.003	0.438 ± 0.003
AugReg-B, 32, 2, T	1.763 ± 0.003	**1.1853** ± 0.0004	71.50 ± 0.16	0.710 ± 0.003	2.330 ± 0.001
AugReg-B, 128, 2, T	1.252 ± 0.006	**0.666** ± 0.006	80.47 ± 0.12	0.663 ± 0.006	1.574 ± 0.006
CLIP-FT-B, 32, 2, T	2.608 ± 0.017	**1.820** ± 0.015	70.43 ± 0.46	1.094 ± 0.017	2.458 ± 0.015
CLIP-FT-B, 128, 2, T	1.1519 ± 0.0005	**0.494** ± 0.001	87.21 ± 0.01	0.5113 ± 0.0005	0.579 ± 0.005
AugReg-B, 64, 4, T	3.347 ± 0.007	**2.763** ± 0.004	31.70 ± 0.62	2.740 ± 0.007	2.335 ± 0.006
CLIP-FT-B, 64, 4, T	3.646 ± 0.002	**2.814** ± 0.002	38.25 ± 0.23	2.997 ± 0.002	1.726 ± 0.007
AugReg-B, 64, 2, R	1.569 ± 0.010	**0.985** ± 0.011	74.71 ± 0.25	0.955 ± 0.010	1.272 ± 0.012
CLIP-FT-B, 64, 2, R	1.435 ± 0.003	**0.755** ± 0.003	83.86 ± 0.25	0.741 ± 0.003	0.4101 ± 0.0004

FSC at 30.41% and 37.04%. Figure 8 shows the same pattern under raw projection: learning helps, and free continuation remains higher for both models.

E.5 Every setting, larger complements, and shifted inputs

Three further groups of runs use the same confirmation, fit, and tuning images and the same recipe: ten Adam epochs, learning rate 3×10^{-3}, selection by natural-teacher KL on the 256-image tuning subset, and three seeds. These runs reuse the confirmation images, so they extend that study rather than forming a second held-out test.

Every setting. FSC is trained in the five settings that the confirmation study above did not cover: AugReg2-B, AugReg-L, DINO-B with its linear readout, MAE-B with its linear probe, and zero-shot CLIP-B. Each setting receives its own fit-only geometric bank (the mean of the fit images' rank-64 future projectors over the entry and the two block outputs of the window, which are block outputs 22 and 23 for the 24-block AugReg-L), the top eight directions of which initialize A after the same small perturbation. The teacher is the setting's own natural output: for zero-shot CLIP-B, the logits of the frozen 80-template text ensemble. Table 21 gives accuracy, CE, and KL with seed SDs. The five simultaneous 95% intervals for Δ_{train} all exclude zero. DINO-B ends within 0.9 points of free continuation and AugReg-L within 0.2. MAE-B improves by 9.9 nats of CE and 14 points of accuracy but remains 43 points below free continuation, consistent with the diagnostic finding that eight directions are too few in this setting, whose readout tolerates little feature error (Appendix D.2).



(a) AugReg-B (b) CLIP-FT-B

Figure 7: **FSC helps across tested budgets.** **(a)** AugReg-B and **(b)** CLIP-FT-B accuracy on the 5,000-image confirmation set, averaged over three source seeds, for the two-block token-norm suffix. Each budget has its own entry projection, excluded support, trained source, and same-entry free reference; lines connect only the evaluated budgets. Clean G_8 nearly coincides with A_0 and is reported in Tables 2, 18, and 19.

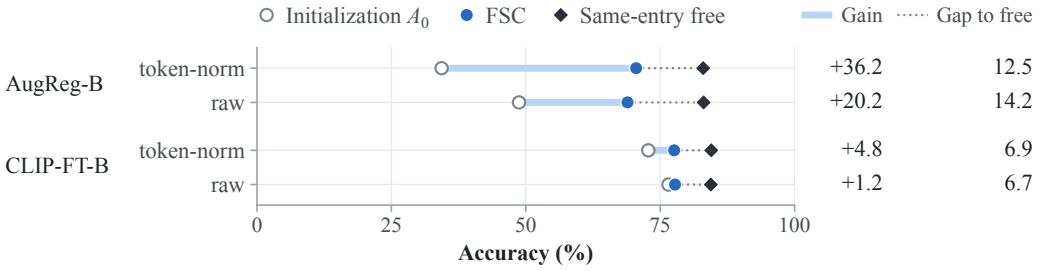


Figure 8: **FSC gains persist under raw projection.** Mean accuracies for two-block, nominal-64 suffixes on the 5,000-image confirmation set. Token-norm rows use the fixed checkpoints; raw rows use sources trained and evaluated under raw projection with the same recipe. Diamonds denote free continuation from the row's own common entry. Segments and columns match Figure 4, and Table 19 lists the values.

Larger complements. With k learned directions the entry keeps $64 - k$ directions, the bank supplies the top k initial directions, and the source has $768 \times k$ parameters. The common entry and free continuation therefore change with k, and the first and third groups of Table 22 are read within rows. Over four blocks, accuracy rises monotonically with $k = 8$, 16, and 32 in both models while the initialization stays near 3% and between 8% and 16%, so the improvement comes from training and the larger bank alone contributes little. Over two blocks, runs with $k = 4$, 16, and 32 extend the main $k = 8$ runs. The gap between FSC and free continuation shrinks from 16.3 points at $k = 4$ to 12.5, 8.9, and 4.7 on AugReg-B, and from 7.5 to 6.9, 6.2, and 4.3 on CLIP-FT-B. Every trained source improves CE and accuracy over its initialization, and the six new training gains have pointwise 95% intervals that exclude zero. A second two-block group holds the entry fixed: every k starts from the entry of $k = 32$, $H^* = T_{\rho,P_{32}}(H_0)$, and only the update basis $[V_{64-k}, C_A(x)]$ changes, so the entry basis and free continuation are the same for all k. FSC then improves at every step, from 63.77% at $k = 4$ to 67.24%, 71.85%, and 75.99% on AugReg-B and from 72.15% to 73.61%, 75.49%, and 78.17% on CLIP-FT-B. Each of the six steps lowers CE by 0.24 to 0.35 nats, with simultaneous 95% intervals that exclude zero. With the reference held fixed, the gap to free continuation narrows from 13.5 points at $k = 8$ to 4.7 at $k = 32$ and from 8.9 to 4.3: replacing entry directions 33 to 56 by learned ones adds 8.8 and 4.6 points. At $k = 8$, starting from the entry of $k = 32$ lowers FSC by 3.3 and 4.0 points, so the first group, whose entry loses information as k grows, understates what a larger complement adds.

Table 21: FSC in the five additional settings ($r = 64$, $k = 8$, last two blocks, token-norm). Seed-dependent entries are mean ± sample SD over three sources; Δ_{train} uses paired CE differences. Accuracy is in %; CE and KL are in nats. Bold highlights FSC accuracy, CE, and KL.

Setting	Free ↑	Entry64 ↑	A_0 ↑	FSC ↑	CE, FSC ↓	KL, FSC ↓	Δ_{train} ↑
AugReg2-B	84.26	74.72	76.93 ± 0.01	**80.09** ± 0.06	**0.923** ± 0.006	**0.330** ± 0.001	1.649 ± 0.006
AugReg-L	84.84	84.06	84.09 ± 0.01	**84.64** ± 0.08	**0.5590** ± 0.0003	**0.0477** ± 0.0003	0.0935 ± 0.0002
DINO-B / linear	71.90	62.42	66.38 ± 0.05	**71.05** ± 0.13	**1.172** ± 0.001	**0.311** ± 0.001	0.215 ± 0.001
MAE-B / linear	57.56	0.34	0.31 ± 0.03	**14.28** ± 0.47	**5.962** ± 0.064	**4.551** ± 0.067	9.867 ± 0.131
CLIP-B / zero-shot	67.80	11.28	28.31 ± 0.13	**55.53** ± 0.23	**1.684** ± 0.008	**0.628** ± 0.005	1.560 ± 0.010

Table 22: Learned-complement size over the last two and four blocks ($r = 64$, token-norm). A_0, FSC accuracy (%), FSC CE, and the paired training gain (nats) are mean ± sample SD over three seeds. In the first and third groups each k has its own entry and free reference; the second starts every k from the entry of $k = 32$ and repeats that row. Bold highlights FSC accuracy and CE.

Model	k	Natural ↑	Free ↑	Entry64 ↑	A_0 ↑	FSC ↑	CE, FSC ↓	Δ_{train} ↑
Last two blocks								
AugReg-B	4	83.98	83.30	51.86	37.34 ± 0.19	**67.03** ± 0.50	**1.697** ± 0.007	1.693 ± 0.006
	8	83.98	83.00	50.64	34.34 ± 0.19	**70.53** ± 0.26	**1.509** ± 0.004	1.962 ± 0.004
	16	83.98	82.78	48.26	42.77 ± 0.08	**73.89** ± 0.18	**1.272** ± 0.002	1.895 ± 0.002
	32	83.98	80.72	36.48	45.31 ± 0.07	**75.99** ± 0.21	**1.089** ± 0.007	2.002 ± 0.009
CLIP-FT-B	4	85.04	84.74	4.30	74.79 ± 0.25	**77.21** ± 0.16	**1.606** ± 0.002	1.173 ± 0.003
	8	85.04	84.50	4.02	72.79 ± 0.19	**77.60** ± 0.24	**1.500** ± 0.002	1.116 ± 0.008
	16	85.04	84.30	2.36	72.09 ± 0.10	**78.08** ± 0.12	**1.396** ± 0.002	1.413 ± 0.006
	32	85.04	82.50	0.40	67.94 ± 0.07	**78.17** ± 0.20	**1.367** ± 0.004	1.970 ± 0.008
Last two blocks, entry of $k = 32$								
AugReg-B	4	83.98	80.72	36.48	33.85 ± 0.08	**63.77** ± 0.40	**1.948** ± 0.002	1.949 ± 0.004
	8	83.98	80.72	36.48	35.61 ± 0.05	**67.24** ± 0.11	**1.712** ± 0.005	2.016 ± 0.005
	16	83.98	80.72	36.48	38.74 ± 0.09	**71.85** ± 0.21	**1.405** ± 0.002	2.206 ± 0.002
	32	83.98	80.72	36.48	45.31 ± 0.07	**75.99** ± 0.21	**1.089** ± 0.007	2.002 ± 0.009
CLIP-FT-B	4	85.04	82.50	0.40	56.35 ± 0.17	**72.15** ± 0.14	**2.230** ± 0.001	2.237 ± 0.008
	8	85.04	82.50	0.40	55.45 ± 0.03	**73.61** ± 0.34	**1.984** ± 0.002	2.475 ± 0.011
	16	85.04	82.50	0.40	60.18 ± 0.27	**75.49** ± 0.12	**1.714** ± 0.008	2.594 ± 0.012
	32	85.04	82.50	0.40	67.94 ± 0.07	**78.17** ± 0.20	**1.367** ± 0.004	1.970 ± 0.008
Last four blocks								
AugReg-B	8	83.98	82.98	0.50	2.69 ± 0.06	**30.41** ± 0.51	**3.347** ± 0.007	2.353 ± 0.009
	16	83.98	82.58	0.46	3.18 ± 0.06	**40.15** ± 0.15	**2.784** ± 0.017	2.736 ± 0.025
	32	83.98	79.62	0.24	3.49 ± 0.05	**48.99** ± 0.38	**2.302** ± 0.003	3.633 ± 0.003
CLIP-FT-B	8	85.04	85.02	0.78	13.91 ± 0.11	**37.04** ± 0.32	**3.646** ± 0.002	1.967 ± 0.010
	16	85.04	84.94	0.70	15.79 ± 0.08	**46.41** ± 0.44	**3.025** ± 0.006	2.601 ± 0.004
	32	85.04	83.62	0.34	8.49 ± 0.09	**55.28** ± 0.68	**2.458** ± 0.003	3.761 ± 0.006

Shifted inputs. The six ViT-B sources selected during development are applied, without retraining or reselection, to the shifted datasets of Appendix D.1: ImageNetV2, ImageNet-Sketch, ImageNet-R, and ImageNet-A for both models and the 75 ImageNet-C conditions of the diagnostic images for AugReg-B, averaged within each image before aggregation. Each shifted image supplies its own entry H^* and entry basis, and the teacher is the unmodified model's prediction on that image. Table 23 reports every arm. FSC improves on both the entry basis and its matched initialization in all nine pairs, and its KL to the teacher lies between 0.97 and 1.42 nats, against 1.9 to 2.9 for the initialization. The share of the entry-basis loss it recovers is largest on ImageNetV2 and smallest on ImageNet-A, where the natural model itself is right on only a quarter of the images.

Table 23: FSC on shifted inputs without retraining. A_0, FSC accuracy (%), and the paired KL difference (nats) are mean ± sample SD over the three selected sources. Accuracy is class-macro. Recovered is the fraction of the entry-basis accuracy gap to free continuation closed by FSC. Bold highlights FSC accuracy and the paired KL change.

Dataset	Free↑	Entry64↑	A_0↑	FSC↑	Recovered↑	KL, FSC − A_0↓
AugReg-B						
ImageNetV2	73.4	42.0	25.7 ± 0.1	$\mathbf{59.5} \pm 0.3$	56%	$\mathbf{-1.886} \pm 0.002$
ImageNet-Sketch	44.4	19.2	9.3 ± 0.1	$\mathbf{27.6} \pm 0.4$	33%	$\mathbf{-1.319} \pm 0.006$
ImageNet-R	43.2	14.5	5.7 ± 0.1	$\mathbf{26.5} \pm 0.5$	42%	$\mathbf{-1.314} \pm 0.011$
ImageNet-A	23.3	4.9	1.3 ± 0.0	$\mathbf{9.2} \pm 0.1$	23%	$\mathbf{-1.211} \pm 0.017$
ImageNet-C	67.1	34.2	19.0 ± 0.1	$\mathbf{49.1} \pm 0.1$	45%	$\mathbf{-1.620} \pm 0.009$
CLIP-FT-B						
ImageNetV2	74.9	2.5	61.6 ± 0.3	$\mathbf{66.7} \pm 0.2$	89%	$\mathbf{-0.970} \pm 0.004$
ImageNet-Sketch	52.4	4.5	36.3 ± 0.1	$\mathbf{40.4} \pm 0.5$	75%	$\mathbf{-0.748} \pm 0.003$
ImageNet-R	50.2	3.0	29.5 ± 0.1	$\mathbf{33.4} \pm 0.1$	65%	$\mathbf{-0.741} \pm 0.010$
ImageNet-A	26.7	1.0	12.0 ± 0.2	$\mathbf{12.9} \pm 0.3$	46%	$\mathbf{-0.751} \pm 0.002$

E.6 Random initialization, training-free sources, and the region of the complement

The next ablations ask what FSC learning depends on and where its directions lie. They use the confirmation, fit, and tuning images and the statistics of Appendix E.5, so they are extensions of the confirmation study. Retraining from the future-informed initialization reproduces the ViT-B sources, selected epochs, and logits behind Table 2.

Random initialization. The source is initialized from natural future activations of the calibration images (Appendix E.2). We retrain both ViT-B sources from a random orthonormal A_0, the orthonormal factor of a $D \times 8$ Gaussian matrix, which uses no data. The recipe, the seeds, and the order of the fit images are unchanged. Random initialization reaches 70.79% on AugReg-B and 77.55% on CLIP-FT-B, against 70.53% and 77.60% for the future-informed initialization (Table 24). The CE of the two trained sources differs by -0.006 nats on AugReg-B (simultaneous 95% interval -0.016 to 0.003) and by $+0.006$ nats on CLIP-FT-B (0.001 to 0.011), at most 0.5% of the improvement over the entry basis. The future-informed source starts far ahead on CLIP-FT-B (72.79% against 4.88%) and behind on AugReg-B (34.34% against 53.33%), and both differences vanish with training. A third start that uses entry information only, ranks 65 to 72 of the eigenbasis of the fit-set mean entry projector, ends at 70.57% and 77.64%, with a CE within 0.01 nats of both others. Measured by the mean squared cosine of the principal angles between two sources, the two initializations start at an overlap of 0.01, the value $8/768$ of two random eight-dimensional subspaces, and end at 0.70 and 0.74. Two randomly initialized seeds overlap by 0.77 and 0.72, and two future-initialized seeds, which start from nearly the same point, by 0.75 and 0.86. Training through the frozen suffix thus finds the complementary directions without future states. An overlap below one at equal accuracy shows that the useful complement is not unique, and a larger complement adds accuracy (Table 22). The tuning curves had essentially plateaued by epoch ten: the largest final-epoch KL drop among random starts was 0.9%.

Training-free sources. Proposition 2 suggests choosing directions by output sensitivity without training, as gradient- and Fisher-weighted compression does (Hsu et al., 2022; Chowdhury et al., 2026; Shihab et al., 2026). We test five such sources. Each is the top eight eigenvectors of a $D \times D$ matrix accumulated over the 6,000 fit images, uses no labels and no optimization, and enters through the construction of Equation (8), exactly as clean G_8 does. Let R collect, over all sites and tokens of an image, the part of the free-continuation updates outside $\mathrm{span}(V_{64}, \mathbf{1})$, let G_y be the gradient of $-\log p_y$ with respect to those updates for a label y drawn from the teacher (eight draws per image), and let $a_y = \langle G_y, R \rangle$ and $\mathrm{sym}(M) = (M + M^\top)/2$. Summed over the fit images, the five matrices are $R^\top R$ (S-PCA, the discarded update energy), $R^\top \mathrm{diag}(w) R$ with w the Fisher trace $\mathbb{E}_y \|G_y\|^2$ of each token (S-Fisher, the analogue of Fisher-weighted factorization), $-\mathrm{sym}(\widetilde{G}^\top R)$ with $\widetilde{G}$ the gradient of the CE of the teacher's top class (S-Grad), $\mathbb{E}_y[(a_y - 1)\,\mathrm{sym}(G_y^\top R)]$ (S-KL2),

Table 24: Source initialization on the confirmation set. Future uses the original start; Random uses a Gaussian orthonormal source; Entry uses ranks 65–72 of the fit-set mean entry projector. Values are mean ± sample SD over three seeds. Δ_{train} is the CE of A_0 minus that of FSC from the same start, and Δ_{init} the CE of FSC minus that of FSC from the Future start, both paired by seed; CE/KL are in nats and accuracies in %.

Model	Start	$A_0\uparrow$	FSC$\uparrow$	CE, FSC$\downarrow$	KL, FSC$\downarrow$	$\Delta_{\text{train}}\uparrow$	$\Delta_{\text{init}}\downarrow$
AugReg-B	Future	34.34 ± 0.19	**70.53** ± 0.26	1.509 ± 0.004	0.925 ± 0.001	—	—
	Random	53.33 ± 0.31	**70.79** ± 0.35	1.502 ± 0.002	0.919 ± 0.002	1.333 ± 0.017	−0.006 ± 0.003
	Entry	51.29 ± 0.05	**70.57** ± 0.37	1.5016 ± 0.0002	0.920 ± 0.001	1.3837 ± 0.0002	−0.007 ± 0.004
CLIP-FT-B	Future	72.79 ± 0.19	**77.60** ± 0.24	1.500 ± 0.002	0.807 ± 0.002	—	—
	Random	4.88 ± 0.35	**77.55** ± 0.13	1.506 ± 0.003	0.811 ± 0.001	5.196 ± 0.036	+0.006 ± 0.001
	Entry	4.49 ± 0.01	**77.64** ± 0.29	1.505 ± 0.003	0.810 ± 0.002	5.232 ± 0.003	+0.005 ± 0.001

and $-\operatorname{sym}(G^\top R)$ with G the gradient of the actual teacher KL and R taken on the rollout confined to V_{56} alone (S-Full1). S-KL2 carries the expansion of Proposition 2 over to the free continuation, with R as the discarded part and the empirical Fisher of the teacher-drawn labels in place of the Hessian. Up to a constant, the teacher KL is then approximated by $-\mathbb{E}_y[a_y] + \frac{1}{2}\mathbb{E}_y[a_y^2]$, restoring orthonormal directions C replaces a_y by $a_y - \operatorname{tr}(C^\top \operatorname{sym}(G_y^\top R)C)$, and the first-order gain is $\operatorname{tr}(C^\top MC)$ for the S-KL2 matrix M. Because a_y sums over all sites before it multiplies $G_y^\top R$, the cross-site terms of the proposition are kept, but R leaves out the entry error and the part of the update residual inside $\operatorname{span}(V_{64}, \mathbf{1})$, which under token-norm comes from rescaling the retained components. S-Full1 linearizes once at the full-strength operating point; under raw projection it gives the linearized decrease of the teacher KL when C is added to V_{56}, and under token-norm it leaves out each token's norm-restoring factor. In each cell we select the source with the lowest teacher KL on the 256-image tuning subset, which also selects the FSC epoch.

Table 25: Training-free sources versus FSC. Each source is the top eight eigendirections of a fit-set energy or sensitivity matrix and is applied through Equation (8). The underlined source has the lowest tuning KL in its cell; accuracy (%) and CE are reported.

	Two blocks, token-norm				Two blocks, raw projection				Four blocks, token-norm			
	AugReg-B		CLIP-FT-B		AugReg-B		CLIP-FT-B		AugReg-B		CLIP-FT-B	
Policy	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓
Entry basis V_{64}	50.64	2.958	4.02	6.833	52.94	2.644	15.70	6.211	0.50	7.135	0.78	6.845
Clean geometric G_8	34.38	3.470	72.88	2.615	48.58	2.860	76.68	1.938	2.70	5.697	14.08	5.609
S-PCA	45.24	2.794	73.38	2.550	55.14	2.303	74.52	2.320	2.22	6.098	11.36	5.708
S-Fisher	44.30	2.698	72.14	2.553	48.98	2.454	73.64	2.258	2.44	6.072	9.24	5.770
S-Grad	31.64	3.215	73.96	2.546	37.76	2.948	75.08	2.280	1.78	6.190	11.62	5.889
S-KL2	43.18	3.056	73.12	2.701	51.70	2.541	74.96	2.327	0.06	7.566	4.08	6.263
S-Full1	41.42	2.894	73.36	2.740	50.38	2.381	74.52	2.324	3.38	5.943	8.46	5.574
FSC (selected A_*)	**70.53**	**1.509**	**77.60**	**1.500**	**68.91**	**1.569**	**77.76**	**1.435**	**30.41**	**3.347**	**37.04**	**3.646**
Free from H^*	83.00	0.614	84.50	0.682	83.08	0.615	84.42	0.694	82.98	0.607	85.02	0.649

None of the five approaches FSC (Table 25). Over two blocks the selected source reaches 44.30% on AugReg-B and 73.96% on CLIP-FT-B, against 70.53% and 77.60% for FSC, and its CE is higher by 1.19 nats (simultaneous 95% interval 1.12 to 1.25) and 1.05 nats (1.01 to 1.08). On AugReg-B all five fall below the entry basis in accuracy. FSC stays far ahead under raw projection, where the first-order derivation applies most directly because no norm is restored (55.14% and 73.64% against 68.91% and 77.76%), and over four blocks (3.38% and 8.46% against 30.41% and 37.04%). Sensitivity does carry information: over two blocks the selected source improves on the geometric source by 0.77 and 0.07 nats. But unweighted update energy does about as well as the weighted forms, and in two of the six cells the geometric source is better than the selected one (Table 26).

Table 26: Selected training-free source compared with FSC and clean G_8. CE differences are in nats; intervals are simultaneous for the primary two-block token-norm rows and pointwise elsewhere. Pred./Meas. compare the leading term of Proposition 2 for the entry basis with its measured KL, both on the confirmation images; overlap reports principal-angle cosine with FSC and between FSC seeds. Bold highlights the CE excess over FSC.

		CE of the selected source minus		KL, entry basis		Overlap,
Model	Selected	vs. FSC CE↓ (95% CI)	vs. G_8 CE↓ (95% CI)	Pred. KL↓	Meas. KL↓	Overlap↑
Two blocks, token-norm						
AugReg-B	S-Fisher	$+\mathbf{1.19}$ $[+1.12, +1.25]$	-0.77 $[-0.83, -0.71]$	0.245	2.369	0.18 / 0.75
CLIP-FT-B	S-Grad	$+\mathbf{1.05}$ $[+1.01, +1.08]$	-0.07 $[-0.11, -0.03]$	0.558	5.659	0.31 / 0.86
Two blocks, raw projection						
AugReg-B	S-PCA	$+\mathbf{0.73}$ $[+0.70, +0.77]$	-0.56 $[-0.59, -0.52]$	0.155	2.055	0.19 / 0.72
CLIP-FT-B	S-Fisher	$+\mathbf{0.82}$ $[+0.80, +0.85]$	$+0.32$ $[+0.29, +0.35]$	0.122	5.079	0.18 / 0.80
Four blocks, token-norm						
AugReg-B	S-Full1	$+\mathbf{2.60}$ $[+2.53, +2.67]$	$+0.25$ $[+0.16, +0.33]$	0.765	6.540	0.14 / 0.82
CLIP-FT-B	S-Full1	$+\mathbf{1.93}$ $[+1.89, +1.97]$	-0.03 $[-0.07, -0.00]$	0.795	5.673	0.23 / 0.85

The local expansion shows why. Evaluated for the entry basis itself on the confirmation images, with its entry error and its actual update residuals, the leading term of Proposition 2 predicts a KL of 0.25 nats on AugReg-B and 0.56 on CLIP-FT-B, whereas the entry basis measures 2.37 and 5.66 nats. Across the six cells the leading term is too small by a factor of 7 to 42. Moving the expansion to the operating point does not help, since S-Full1 is selected only over four blocks and stays far from FSC there. Restoring a direction is a finite step that a single linearization does not capture, which is why FSC optimizes the actual KL at full strength (Section 4). The selected sources overlap the learned subspace by 0.14 to 0.31, against 0.72 to 0.86 between FSC seeds. Together with the random initialization, these runs locate FSC's gain: it comes neither from future states nor from a one-shot sensitivity estimate, but from optimizing the prediction loss through the frozen suffix.

The learned complement and the complete entry. Equation (8) excludes $\mathrm{span}(V_{64}, \mathbf{1})$, so the complement could still be drawn from the entry's own trailing directions, ranks 65 to 197. It draws on them no more than a random subspace does. For each confirmation image we measure the share of the applied complement $C_A(x)$ that lies outside the complete entry support $\mathrm{span}(V_{197}, \mathbf{1})$. For the FSC sources this share is 81.5% on AugReg-B and 80.3% on CLIP-FT-B, against 81.1% for a random subspace of the same dimension ($570/703$). The geometric initialization starts at 74.2% and 73.2%, the sources trained from the random and the entry-only start end at 81.5% and 80.4%, and the four-block sources with $k = 8$, 16, and 32 lie between 81.0% and 81.3%. The learned directions thus show no preference for the entry's trailing directions. The diagnostics agree on the functional side: on the 2,000 diagnostic images, 56 entry directions with eight future-derived ones give a higher accuracy than the complete 197-column entry basis in seven of the eight settings, all but DINO-B, although that basis contains every entry direction and starts from a lossless entry (Table 8).

Confining the complement. A third group of runs tests this directly for the learned directions. For each image let $T(x)$ be an orthonormal basis of the part of the complete entry support $\mathrm{span}(V_{197}, \mathbf{1})$ that lies outside $\mathrm{span}(V_{64}, \mathbf{1})$, 133 directions. The *outside* arm uses $C_A(x) = \mathrm{orth}((I - U_r U_r^\top - TT^\top)A)$ in Equation (8), which confines the complement to the 570 dimensions outside the complete entry support, and the *inside* arm uses $C_A(x) = \mathrm{orth}(TT^\top A)$, which confines it to the entry's trailing directions. The entry H^*, budget, initialization A_0, recipe, fit-image order, and selection rule are unchanged. Confined to the outside, FSC reaches 68.69% on AugReg-B and 77.41% on CLIP-FT-B, against 70.53% and 77.60%, and keeps 93% and 98% of its CE gain over the entry basis (Table 27). Confined to the entry's trailing directions it reaches 60.13% and 34.87% and keeps 49% and 27%. The CE of the outside arm exceeds that of FSC by 0.099 nats (simultaneous 95% interval 0.084 to 0.114) and 0.080 nats (0.069 to 0.092), so the trailing directions contribute, but little, and the inside arm is worse than the outside arm by 0.645 nats (0.609 to 0.681) and 3.817 nats (3.763 to 3.870). Applied to the trained FSC sources without retraining, the two confinements split the gain the same way (Table 27): outside the complete entry support, the sources reach 68.73% and 77.35% and keep 92.1% and 97.4% of it; inside the trailing directions, they reach 57.37% and 19.45% and keep 30.5%

and 14.8%. Learning helps inside the entry as well: the entry basis is itself V_{56} completed by the next eight entry directions in order of energy, and the learned inside arm improves on it by 9.5 and 30.8 points. Yet the entry as a whole is not enough. Applied from the same H^*, all 197 entry directions reach 65.86% and 41.12%, below the eight learned directions from outside. The same ordering holds under raw projection and over four blocks: in all six cells the outside arm keeps more than 92% of the gain and the inside arm less than 49%, and the eight outside directions exceed the complete entry basis by 2.8 to 36.3 points.

Table 27: Location of the learned complement. Outside is confined outside complete entry support; Inside uses trailing entry directions. The second group confines the trained FSC sources without retraining. FSC, Outside, and Inside accuracies (%) are mean ± sample SD over three seeds. R_{out} and R_{in} are the shares (%) of FSC's seed-mean CE gain retained by each confined arm, with pointwise 95% CIs. Bold highlights Outside accuracy and its retained gain.

Model	V_{64} ↑	V_{197} ↑	FSC ↑	Outside ↑	Inside ↑	Free ↑	R_{out} ↑ (95% CI)	R_{in} ↑ (95% CI)
Two blocks, token-norm								
AugReg-B	50.64	65.86	70.53 ± 0.26	$\mathbf{68.69} \pm 0.41$	60.13 ± 0.20	83.00	**93.17** [92.42, 93.92]	48.66 [47.54, 49.76]
CLIP-FT-B	4.02	41.12	77.60 ± 0.24	$\mathbf{77.41} \pm 0.10$	34.87 ± 0.22	84.50	**98.49** [98.33, 98.66]	26.92 [26.58, 27.26]
Two blocks, token-norm, without retraining								
AugReg-B	50.64	65.86	70.53 ± 0.26	$\mathbf{68.73} \pm 0.09$	57.37 ± 0.19	83.00	**92.07** [91.55, 92.59]	30.48 [29.52, 31.44]
CLIP-FT-B	4.02	41.12	77.60 ± 0.24	$\mathbf{77.35} \pm 0.15$	19.45 ± 0.02	84.50	**97.36** [97.19, 97.53]	14.82 [14.56, 15.09]
Two blocks, raw projection								
AugReg-B	52.94	63.86	68.91 ± 0.28	$\mathbf{67.09} \pm 0.10$	58.76 ± 0.03	83.08	**92.50** [91.59, 93.39]	39.20 [38.11, 40.28]
CLIP-FT-B	15.70	59.20	77.76 ± 0.16	$\mathbf{77.78} \pm 0.24$	65.31 ± 0.15	84.42	**98.95** [98.78, 99.12]	43.69 [43.09, 44.28]
Four blocks, token-norm								
AugReg-B	0.50	5.90	30.41 ± 0.51	$\mathbf{27.09} \pm 0.30$	5.25 ± 0.08	82.98	**93.83** [93.25, 94.42]	36.75 [35.98, 37.52]
CLIP-FT-B	0.78	22.98	37.04 ± 0.32	$\mathbf{32.06} \pm 0.20$	13.40 ± 0.14	85.02	**92.70** [92.24, 93.14]	41.34 [40.59, 42.09]

Widening the complement. Equation (8) also keeps the complement out of entry directions 57 to 64, which the entry basis holds beyond V_{56}. Cleaning the source only against $\mathrm{span}(V_{56}, \mathbf{1})$ admits them. Trained this way from the same A_0 with the same recipe, the widened source reaches 70.49% on AugReg-B and 77.34% on CLIP-FT-B, against 70.53% and 77.60% for FSC, and attains 100.5% and 99.1% of FSC's CE gain over the entry basis (pointwise 95% intervals 100.3% to 100.8% and 99.0% to 99.2%). The share of its complement that lies in those eight directions is 1.2% and 1.8%, against $8/711 = 1.1\%$ for a random subspace and 3.2% at initialization. Excluding them therefore costs nothing, and it guarantees that every learned direction lies outside the entry basis.

E.7 Consistency across source initializations

All 33 selected sources improve both CE and accuracy relative to their saved initialization on the confirmation set. Table 2 summarizes the three selected sources of each ViT-B by their mean and standard deviation, and Table 28 lists those six pairs individually, together with the 27 extension runs and their selected epochs. The 33 pairs share the same 5,000 images, so the table shows consistency across source initializations.

Appendix H.2 describes the supplementary code, which re-evaluates the six sources of Table 2 and trains a source with the same recipe.

F Development-set comparisons and source attribution

The construction of a source matters before any learning, and the comparisons in this appendix measure how much. They use the 2,000-image development set, also used by the matched-source analysis below. The predicted-future, G_{64}, and privileged-future policies were evaluated only on this development set.

Table 28: Source-specific confirmation outcomes for the 33 checkpoint-initialization pairs. The first six pairs are the development-selected checkpoints; the remaining 27 are fixed-recipe extensions. All pairs are evaluated on the same 5,000 images. Bold highlights FSC endpoints.

Configuration	Seed	Epoch	A_0 CE↓	FSC CE↓	A_0 acc.↑	FSC acc.↑
AugReg-B, 64, 2, T	0	8	3.469	**1.507**	34.12	**70.26**
	1	10	3.471	**1.506**	34.44	**70.78**
	2	9	3.470	**1.513**	34.46	**70.54**
CLIP-FT-B, 64, 2, T	0	7	2.623	**1.503**	72.64	**77.84**
	1	10	2.620	**1.499**	72.74	**77.60**
	2	10	2.605	**1.499**	73.00	**77.36**
AugReg-S, 64, 2, T	0	8	1.678	**1.225**	62.18	**71.16**
	1	6	1.687	**1.230**	61.80	**71.02**
	2	7	1.688	**1.231**	61.80	**70.94**
AugReg-B, 32, 2, T	0	10	4.118	**1.763**	28.34	**66.88**
	1	9	4.119	**1.760**	28.42	**67.10**
	2	7	4.119	**1.766**	28.26	**66.76**
AugReg-B, 128, 2, T	0	8	2.846	**1.245**	48.24	**74.18**
	1	10	2.848	**1.253**	48.16	**74.22**
	2	6	2.847	**1.257**	48.52	**74.32**
CLIP-FT-B, 32, 2, T	0	10	5.345	**2.604**	43.80	**67.48**
	1	9	5.337	**2.626**	44.18	**66.52**
	2	10	5.326	**2.593**	44.14	**66.86**
CLIP-FT-B, 128, 2, T	0	10	1.863	**1.151**	78.20	**80.48**
	1	10	1.858	**1.152**	78.48	**80.42**
	2	8	1.854	**1.152**	78.42	**80.48**
AugReg-B, 64, 4, T	0	8	5.697	**3.351**	2.76	**30.92**
	1	8	5.702	**3.351**	2.64	**29.90**
	2	10	5.702	**3.338**	2.68	**30.40**
CLIP-FT-B, 64, 4, T	0	6	5.615	**3.648**	13.78	**36.92**
	1	9	5.603	**3.645**	13.96	**37.40**
	2	10	5.623	**3.645**	13.98	**36.80**
AugReg-B, 64, 2, R	0	8	2.860	**1.562**	48.68	**69.18**
	1	5	2.858	**1.581**	48.70	**68.62**
	2	9	2.860	**1.565**	48.82	**68.94**
CLIP-FT-B, 64, 2, R	0	5	1.942	**1.438**	76.42	**77.58**
	1	10	1.936	**1.434**	76.50	**77.84**
	2	9	1.936	**1.433**	76.68	**77.86**

F.1 Complement construction can dominate the untrained baseline

Table 29 compares constrained policies at the same entry, window, and nominal budget, alongside natural and same-entry free references. A ridge predictor estimates a future projector from the current entry's token mean. Shared geometric alternatives use directions of the mean fit-set future projector. The G_{64} recipe residualizes its top 64 candidate directions and selects eight, whereas the clean G_8 recipe selects the top eight *before* residualization, then uses exactly the FSC source construction. These are different recipes despite the same final budget. Adding the small initialization perturbation to clean G_8 gives A_0, and A_* is the trained source selected from it.

FSC reaches 70.28% accuracy on AugReg-B and 76.62% on CLIP-FT, compared with 58.50% and 47.35% for the tested input-conditioned predictor. Clean G_8 already reaches 71.85% on CLIP-FT, exceeding the input-conditioned predictor before any source learning. On AugReg, it yields only 36.45%. This contrast makes the construction-matched baseline decisive: a large difference between FSC and G_{64} can combine a better source recipe with a learning gain. The next comparison separates those contributions.

Table 29: **Source recipes on the development set.** Accuracy (%), CE change from the natural model, and teacher KL. All constrained arms use the same entry H^* and 64 selected columns; Future is the privileged diagnostic. A_0 and FSC are mean ± sample SD over three seeds.

	AugReg ViT-B			CLIP-FT ViT-B		
Policy	Acc. ↑	ΔCE ↓	KL ↓	Acc. ↑	ΔCE ↓	KL ↓
Natural model	84.05	0.000	0.000	84.55	0.000	0.000
Free from H^*	83.55	0.022	0.029	84.15	0.048	0.028
Entry64	52.85	2.331	2.311	3.65	6.200	5.661
Shared geometric G_{64}	54.00	2.370	2.344	22.35	5.236	4.765
Clean geometric G_8	36.45	2.853	2.814	71.85	2.007	1.788
Predicted future	58.50	1.639	1.628	47.35	4.215	3.828
Matched initialization A_0	36.30 ± 0.09	2.8539 ± 0.0002	2.8147 ± 0.0001	71.88 ± 0.03	2.008 ± 0.009	1.788 ± 0.008
FSC (selected A_*)	**70.28** ± 0.37	**0.909** ± 0.003	**0.906** ± 0.003	**76.62** ± 0.19	**0.883** ± 0.001	**0.812** ± 0.001
Privileged future	80.35	0.242	0.245	82.55	0.940	0.819

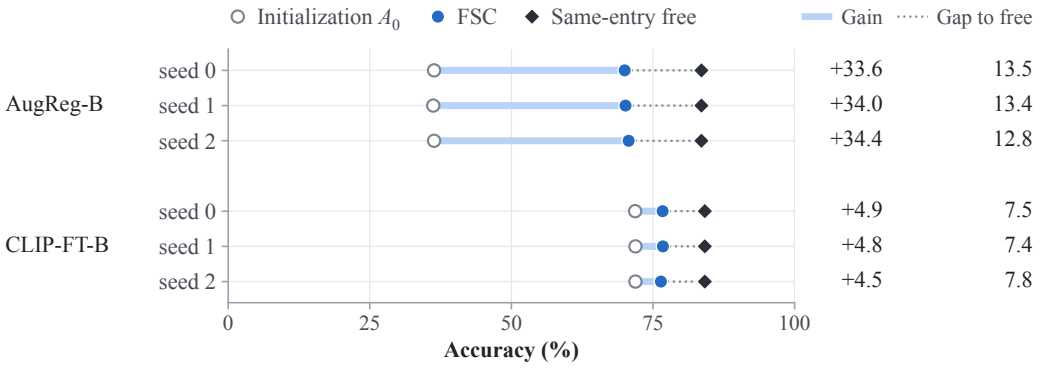


Figure 9: **Training improves the same source parameterization.** On the 2,000-image development set, each shaded segment links an initialization to the checkpoint selected from it; dotted segments extend to same-entry free continuation (diamonds). Horizontal position is accuracy, and the right-hand columns report both segment lengths in percentage points. Every model-seed pair improves accuracy, CE, KL, and teacher agreement, isolating the effect of learning within a fixed recipe.

F.2 Separating construction, initialization, and training

Each epoch-zero source is paired with the checkpoint selected from it, with source shape, application rule, common entry, and evaluation images held fixed. Three initializations and clean G_8 are evaluated per model without new training or selection. Appendix E repeats this matched comparison on held-out images.

Averaging metrics within image across the three seeds, accuracy rises from 36.30% to 70.28% on AugReg and from 71.88% to 76.62% on CLIP-FT (Figure 9). CE decreases by 1.945 nats on AugReg (simultaneous 95% interval 1.880 to 2.011) and by 1.125 nats on CLIP-FT (1.081 to 1.168). These intervals use 10,000 synchronized class-cluster bootstrap draws in a separate two-member CE family, for the given checkpoints, tuning choices, and seeds.

For mean CE ℓ, the difference between FSC and the 64-column geometric recipe separates exactly into three terms:

$$\ell(A_*) - \ell(G_{64}) = \underbrace{\ell(G_8) - \ell(G_{64})}_{\text{source construction}} + \underbrace{\ell(A_0) - \ell(G_8)}_{\text{initialization}} + \underbrace{\ell(A_*) - \ell(A_0)}_{\text{training and selection}}. \tag{F.1}$$

The three terms are $(+0.483, +0.0006, -1.945)$ nats for AugReg and $(-3.229, +0.0004, -1.125)$ for CLIP-FT. The construction term dominates on CLIP-FT and even has the opposite sign on

AugReg, whereas the training term improves both models. The decomposition is an exact additive identity in CE for these three recipes.

G Operation count of a folded suffix

The experiments constrain residual-update directions while still executing the full-width model. Table 30 asks what a suffix could save if it instead computed inside the constrained support while preserving predictions. A full-width ViT-B block costs $12ND^2 + 2N^2D$ multiply-adds ($N = 197$, $D = 768$, 12 heads, MLP width $4D$). A folded block with an s-dimensional residual stream costs $12NDs + 2N^2D$, plus $12D^2s$ per image and block to fold the branch weights into the image-specific basis. Attention scores keep their full N^2D cost.

Raw projection has $s = r$ and token-norm has $s = r + 1$. The count includes the entry SVD, FSC basis construction, support exclusion in Equation (8), and readout reconstruction. Token-norm also needs each update token's original mean and centered norm; the ideal column treats these statistics as already available, while the last column charges two full-width maps per block to compute them. At $r = 64$, the folded suffix costs about 45% of full width in the ideal case and about 70% when token-norm statistics are computed. The counts are per image and do not model batching overhead from image-specific folded weights.

Table 30: Analytic multiply-add counts for a ViT-B suffix computed inside the constrained support. Percentages are savings relative to full width. The folded counts include entry SVD, FSC basis construction, image-specific weight folding, and full attention-score cost; token-norm is shown with and without charging full-width statistics.

r	Blocks	Full width (GMADD) ↓	Raw savings ↑	Token-norm savings ↑	Savings with stats ↑
32	2	2.91	75.1%	74.4%	42.1%
64	2	2.91	55.3%	54.6%	29.9%
128	2	2.91	15.8%	15.0%	5.3%
32	4	5.82	75.7%	75.0%	42.8%
64	4	5.82	56.0%	55.3%	30.6%
128	4	5.82	16.7%	16.0%	6.2%

Removing the window. The simplest way to save the window's cost is to skip it: every residual update of the window is set to zero, so the natural entry passes directly to the readout. FSC is more accurate than this removal in all 20 configurations of Table 31, by 3.8 to 65.3 points, and its CE is lower in each, with every pointwise 95% interval excluding zero. Removal alone reaches only 8.98% on AugReg-B, 16.04% on AugReg-S, 0.96% for zero-shot CLIP-B, 0.48% on MAE-B, and below 2% over four blocks; the loss is smaller in CLIP-FT-B (62.82%), AugReg2-B (65.28%), DINO-B (63.34%), and AugReg-L (80.84%).

H Implementation and reproducibility

H.1 Test-time application and information access

FSC selects a space once for each input and uses it throughout the chosen suffix. The frozen prefix produces the entry, V_r is computed, and the shared entry is $H^* = T_{\rho,P_{r-8}}(H_0)$. Equation (8) removes the full $\mathrm{span}(V_r, \mathbf{1})$ from the learned source and appends the resulting eight directions to V_{r-8}. Each subsequent branch is evaluated on the current constrained state, and its update is transformed before residual addition. The basis remains fixed for all four or eight additions.

Natural future states on calibration images initialize the shared source, and natural predictions supply the training target. The initialization can be replaced by a random one (Appendix E.6). At test time, constructing the complement requires only the learned source and the current entry basis, unlike the privileged-future diagnostic in Table 32.

Table 31: Removing the constrained window. Accuracy (%) on the 5,000 confirmation images when every residual update of the window is set to zero, starting from the natural entry, beside the entry basis and FSC (seed mean) of the same configuration. The last two columns give FSC minus removal in accuracy points and removal minus FSC in CE (nats), with pointwise 95% class-bootstrap intervals.

Setting	r	Natural ↑	Entry ↑	FSC ↑	Removed ↑	Acc. gain ↑	CE gain ↑
Two blocks, token-norm							
AugReg-B	64	83.98	50.64	**70.53**	8.98	61.5 [59.7, 63.4]	4.506 [4.456, 4.557]
CLIP-FT-B	64	85.04	4.02	**77.60**	62.82	14.8 [13.4, 16.2]	2.619 [2.588, 2.649]
AugReg2-B	64	84.32	74.72	**80.09**	65.28	14.8 [13.5, 16.1]	3.101 [3.061, 3.142]
AugReg-S	64	80.56	56.40	**71.04**	16.04	55.0 [53.2, 56.9]	4.789 [4.740, 4.837]
AugReg-L	64	85.00	84.06	**84.64**	80.84	3.8 [3.0, 4.6]	0.435 [0.409, 0.461]
DINO-B / linear	64	76.64	62.42	**71.05**	63.34	7.7 [6.3, 9.1]	0.451 [0.409, 0.493]
MAE-B / linear	64	66.74	0.34	**14.28**	0.48	13.8 [12.7, 14.9]	28.375 [27.587, 29.151]
CLIP-B / zero-shot	64	67.96	11.28	**55.53**	0.96	54.6 [52.8, 56.4]	7.042 [6.827, 7.260]
AugReg-B	32	83.98	26.86	**66.91**	8.98	57.9 [56.1, 59.8]	4.252 [4.201, 4.303]
AugReg-B	128	83.98	61.38	**74.24**	8.98	65.3 [63.5, 67.1]	4.763 [4.713, 4.812]
CLIP-FT-B	32	85.04	2.98	**66.95**	62.82	4.1 [2.4, 5.9]	1.511 [1.469, 1.553]
CLIP-FT-B	128	85.04	20.34	**80.46**	62.82	17.6 [16.2, 19.1]	2.967 [2.934, 3.000]
Two blocks, raw projection							
AugReg-B	64	83.98	52.94	**68.91**	8.98	59.9 [58.1, 61.7]	4.446 [4.395, 4.496]
CLIP-FT-B	64	85.04	15.70	**77.76**	62.82	14.9 [13.6, 16.3]	2.684 [2.653, 2.715]
Four blocks, token-norm							
AugReg-B, $k = 8$	64	83.98	0.50	**30.41**	1.86	28.5 [27.1, 30.1]	2.711 [2.659, 2.764]
AugReg-B, $k = 16$	64	83.98	0.46	**40.15**	1.86	38.3 [36.7, 39.9]	3.274 [3.219, 3.330]
AugReg-B, $k = 32$	64	83.98	0.24	**48.99**	1.86	47.1 [45.5, 48.8]	3.756 [3.696, 3.817]
CLIP-FT-B, $k = 8$	64	85.04	0.78	**37.04**	1.14	35.9 [34.3, 37.5]	2.886 [2.844, 2.929]
CLIP-FT-B, $k = 16$	64	85.04	0.70	**46.41**	1.14	45.3 [43.6, 47.0]	3.507 [3.463, 3.553]
CLIP-FT-B, $k = 32$	64	85.04	0.34	**55.28**	1.14	54.1 [52.5, 55.9]	4.075 [4.026, 4.124]

H.2 Data and reproduction scope

The supplementary code contains a small library for the interventions of Section 3, the common entry, the completed basis of Equation (8) with its confined and widened variants, and the eight frozen settings, together with scripts for each experiment. They measure the subspace distances of Section 3 and rerun the view control of Appendix B.4; evaluate the supplied sources of each configuration and ablation, together with clean G_k and the reference arms, on the confirmation images or a shifted dataset; evaluate the removed window of Appendix G; rerun the complete-entry and common-entry controls of Section 5 and Appendix D for any setting, window, projection rule, dataset, and number of replaced directions, with their entry, update-energy, readout, and per-token rescaling diagnostics; compare the source recipes of Appendix F, including the ridge predictor; train a source with the recipe of Table 7 from each initialization and in each region of Appendix E.6; build the training-free sources and evaluate the local expansion of Proposition 2; compute class-bootstrap intervals from per-image outputs; count the operations of Appendix G; and check Propositions 1 and 2 on constructed systems.

Image manifests identify every image by file hash, the checkpoints and official implementations are pinned by revision and hash, and the MAE-B probe is included. Run on the public checkpoints, the scripts reproduce Figures 2–9 and Tables 1–4, 6, and 8–31, including all their bootstrap intervals, and retraining reproduces the selected epoch and the saved source in each configuration we reran.

H.3 Frozen encoders, readouts, and preprocessing

The five pretrained classifier settings use the `timm` checkpoints listed in Table 33, including their original heads (Wightman, 2019). The remaining three settings use official DINO, MAE, and OpenAI CLIP implementations (Caron et al., 2021; He et al., 2022; Radford et al., 2021), with the following readouts and preprocessing fixed throughout the interventions. The supplementary code pins the revision of every checkpoint and repository.

Table 32: All arms of the study, with what each applies to the window's updates and what it uses to select directions for a test input. All arms use the current entry, and none uses test labels. Complete-entry arms project the entry onto 197 columns; common-entry arms start from H^* with nominal budget r (64 unless noted). "Future" denotes natural states after the entry.

Arm	Window updates	Calibration information	Test future
Complete entry (Section 5)			
Entry-only	Unrestricted	None	No
Persistent-entry	Complete entry basis	None	No
Entry-only-future	Unrestricted	None	Yes
Persistent-future	Future basis	None	Yes
Common entry H^* (Sections 5.2 and 6)			
Free	Unrestricted	None	No
Entry-r	V_r	None	No
Random complement	V_{r-8} and random directions	Fixed random source	No
Privileged future	V_{r-8} and future directions	None needed	Yes
G_{64} / clean G_8	V_{r-8} and shared directions	Fit-set future projectors	No
Predicted future	V_{r-8} and predicted directions	Fit current features and future projectors	No
FSC	$[V_{r-k}, C_A(x)]$	Fit future projectors for initialization, fit teacher KL for training, tune KL for selection	No
FSC, random start	$[V_{r-k}, C_A(x)]$	Fit teacher KL for training, tune KL for selection	No
Training-free sources	$[V_{r-8}, C(x)]$	Fit free-continuation updates and teacher-sampled gradients, tune KL for selection	No
Natural entry (Appendix G)			
Removed window	Set to zero	None	No

Table 33: Frozen models and public checkpoints. The first five entries give `timm` model names and pretrained tags; the remaining entries identify the official implementation or weight file.

Setting	Checkpoint
AugReg-B	`vit_base_patch16_224.augreg_in21k_ft_in1k`
AugReg2-B	`vit_base_patch16_224.augreg2_in21k_ft_in1k`
CLIP-FT-B	`vit_base_patch16_clip_224.openai_ft_in1k`
AugReg-S	`vit_small_patch16_224.augreg_in21k_ft_in1k`
AugReg-L	`vit_large_patch16_224.augreg_in21k_ft_in1k`
DINO-B / linear	Official DINO `vit_base`, patch 16
MAE-B / linear	`mae_pretrain_vit_base.pth`
CLIP-B / zero-shot	OpenAI CLIP `ViT-B-16.pt`

DINO-B with a linear readout. The encoder is the official `vit_base(patch_size=16, num_classes=0)`. The 1,536-dimensional input to the official linear head interleaves the final normalized CLS token and patch mean channel by channel: $(\mathrm{CLS}_0, \mathrm{mean}_0, \mathrm{CLS}_1, \mathrm{mean}_1, \ldots)$. It uses one final block, not a concatenation of four blocks. Images are resized to 256 pixels on the shorter side and center-cropped to 224×224 with bicubic interpolation, and normalization uses mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$.

MAE-B with a fixed linear probe. The encoder is the official `vit_base_patch16`, run with `global_pool=False`, no masking, and the final normalized CLS token. The probe is non-affine BatchNorm with $\epsilon = 10^{-6}$ followed by a 1,000-way linear layer. It is trained once on 1,281,167 ImageNet training images with the encoder frozen, official LARS, 90 epochs, global batch 4,096, local batch 512, zero weight decay, and seed 0. The base learning rate 0.1 is scaled by global batch

size divided by 256, giving a peak of 1.6 after 10 warm-up epochs, followed by cosine decay. The final epoch is fixed without validation-based selection. Its training augmentation uses the official MAE random-resized crop with scale $[0.08, 1.0]$ and horizontal flip, and evaluation uses the DINO normalization and resize/crop settings above. The probe is frozen during every diagnostic, and its training is separate from FSC source learning. Its BatchNorm layer then runs in inference mode with the running statistics accumulated during probe training, so an image's prediction does not depend on its batch. Appendix D.2 describes what this standardization does to the intervention results.

CLIP-B with a zero-shot readout. The official OpenAI `ViT-B/16` checkpoint is loaded without JIT. Each ImageNet class uses the official 80-template prompt ensemble. Text embeddings are normalized per template, averaged, and normalized again, and normalized image embeddings use the checkpoint logit scale, which is 100.0. Images are resized to 224 pixels on the shorter side and center-cropped to 224×224 with bicubic interpolation. The normalization mean is $(0.48145466, 0.45782750, 0.40821073)$ and standard deviation $(0.26862954, 0.26130258, 0.27577711)$. The class names are those of the official repository.