SEMANTIC CAPABILITY ACQUISITION AND SPECIALIZATION DURING VISION-LANGUAGE MODEL FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning vision-language models (VLMs) is typically evaluated at a single downstream checkpoint, obscuring whether a semantic capability was never acquired or emerged earlier and later declined during specialization. We ask how semantic capabilities are acquired, when they peak, how well they transfer, and what remains at deployment. We study these dynamics as a *semantic capability trajectory*, tracking identity- and attribute-based capabilities over training.

We formulate a trajectory-based framework that separates capability acquisition, capability-specific optima, and later specialization, and introduce Structured Semantic Routing (SSR) to study how the representation of supervision shapes what is acquired. Across six pretrained backbones spanning DFN, MetaCLIP, and OpenAI CLIP, we show that fine-tuning can acquire semantic capability beyond the pretrained state, including gains observed on held-out evaluations. Unstructured name-and-attribute supervision produces strong name-and-attribute retrieval with comparatively weak name-free attribute-profile retrieval, whereas SSR yields substantially stronger name-free attribute-profile retrieval and is further strengthened by stochastic name-branch dropout. Different capabilities can peak at different stages, so a checkpoint selected by target class-name retrieval need not coincide with a transferable semantic optimum. Continued optimization can therefore preserve strong target class-name retrieval while reducing previously acquired transferable semantic capability. In a representative diagnostic study, this late specialization is consistent with reduced cross-modal semantic accessibility while substantial image-only class structure remains available.

1 INTRODUCTION

Vision–language models (VLMs) are commonly adapted by fine-tuning a pretrained model and evaluating a single resulting checkpoint. This endpoint view cannot distinguish a semantic capability that was never acquired from one that emerged during training and later declined, or remained represented but became less accessible through the cross-modal interface. Prior work has shown that robustness can change substantially over fine-tuning and depends on the adaptation strategy (Andreassen et al., 2021; Wortsman et al., 2022; Goyal et al., 2023). We ask a capability-specific question: *what semantic capabilities are acquired during fine-tuning, when do they peak, how well do they transfer, and what remains at deployment?*

This question is especially important when class names provide strong identity anchors. Richer language, fine-grained captions, and structured alignment can improve VLMs (Menon & Vondrick, 2022; Xie et al., 2025b;a), yet strong name-and-attribute retrieval does not necessarily indicate strong use of attributes. WaffleCLIP shows that performance gains from adding descriptive text can persist even when the added text is only weakly semantic (Roth et al., 2023), while controlled tests that change one attribute or relation at a time show that VLMs may rely less on those details than standard retrieval scores suggest (Yuksekgonul et al., 2022; Bader et al., 2025). A zero-shot CUB analysis likewise shows that simply adding descriptive text to a class-name prompt can reduce retrieval (Appendix B.8), further motivating the broader question: *which semantic capability is actually acquired?*

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.

We study fine-tuning as a *semantic capability trajectory*. We track four distinct probes: class-name retrieval (P1), name-and-attribute retrieval (P2), name-free attribute-profile retrieval (P3), and Counterfactual Attribute-Profile Preference (CAPP). For each capability, we distinguish the pretrained state, acquisition beyond that state, its capability-specific optimum, and what remains at the checkpoint selected independently by target-domain class-name retrieval. This separates capability acquisition from subsequent specialization and allows different semantic capabilities to reach their optima at different stages of the same fine-tuning run.

To study how supervision affects these trajectories, we introduce *Structured Semantic Routing* (SSR), which represents class identity and descriptive attributes through separate semantic routes during fine-tuning. Across six pretrained backbones spanning DFN (Fang et al., 2024), MetaCLIP (Xu et al., 2024), and OpenAI CLIP (Radford et al., 2021), we find that fine-tuning can acquire semantic capability beyond the pretrained state, including gains observed on held-out evaluations. However, what is acquired depends strongly on how supervision is represented: unstructured name-and-attribute supervision can produce strong name-and-attribute retrieval despite comparatively weak name-free attribute-profile retrieval, whereas SSR produces substantially stronger name-free capability, further strengthened by stochastic name-branch dropout. Different capabilities can peak at different stages, so continued optimization may preserve strong target class-name retrieval while reducing transferable semantic capability acquired earlier.

Finally, we examine what changes between an earlier transferable optimum and a later specialized state. Across a broad trajectory grid, greater peak-to-deployment backbone movement is associated with larger held-out performance decreases in both class-name retrieval (P1) and name-free attribute-profile retrieval (P3), indicating that this association is not specific to the name-free semantic probe. In a representative P3 transition, similarity to both the correct and competing semantic anchors increases, but more strongly for the competitor, reducing the correct-minus-wrong margin while substantial image-only class structure remains accessible. We interpret this representative pattern as reduced cross-modal semantic accessibility.

Our contributions are:

- We formulate VLM fine-tuning as a semantic capability trajectory and measure capability acquisition, capability-specific optima, deployment-time capability, and later specialization.
- We introduce Structured Semantic Routing (SSR) and show across six diverse pretrained backbones that supervision content and semantic routing can produce substantially different acquired capability profiles.
- We show that transferable semantic capability can peak before a target-selected deployment checkpoint, and we provide representative diagnostic evidence that late degradation can reflect reduced cross-modal semantic accessibility while substantial visual semantic structure remains accessible.

2 Related Work

Fine-tuning trajectories and preservation. Fine-tuning can improve target performance while changing robustness outside the adaptation distribution. Andreassen et al. (2021) explicitly track effective OOD robustness throughout fine-tuning, showing that inherited robustness can be transient; WiSE-FT (Wortsman et al., 2022) recovers robustness by interpolating pretrained and fine-tuned weights, FLYP (Goyal et al., 2023) changes downstream behavior by preserving the contrastive pretraining objective, and LiT (Zhai et al., 2022) learns cross-modal alignment while one modality remains fixed. These works establish that optimization stage and adaptation strategy matter. Our focus is capability-specific: we ask whether a semantic capability is inherited or newly acquired, when it peaks, and whether it remains transferable at deployment.

Descriptive supervision and semantic diagnostics. Classification-by-description methods enrich class names with visual descriptors (Menon & Vondrick, 2022), while WaffleCLIP shows that such prompt gains need not isolate useful descriptive semantics (Roth et al., 2023). Name-free descriptive recognition has also been studied directly (Baron et al., 2024), including with large VLMs (Atabuzzaman et al., 2025); FG-CLIP and FG-CLIP 2 add long captions, regional alignment, hard negatives, and stronger fine-grained textual discrimination during training (Xie et al., 2025b;a). Other

diagnostics expose failures hidden by aggregate accuracy: ARO/NegCLIP and VL-CheckList test compositional or attribute sensitivity (Yuksekgonul et al., 2022; Zhao et al., 2022), FG-OVD uses attribute-modified alternatives (Bianchi et al., 2024), and SUB shows that concept predictions can follow class-correlated expectations rather than modified visual evidence (Bader et al., 2025). Our novelty is not name removal or counterfactual testing itself, but using distinct probes longitudinally to measure acquisition, peak timing, and retention.

Cross-modal accessibility and grounding. Poor multimodal output need not imply absent visual information. Hidden in Plain Sight finds information readable from visual features that the full VLM fails to exploit (Fu et al., 2025), while VLMs Need Words documents strong reliance on nameable semantic anchors (Shahgir et al., 2026). Modality-gap work further cautions that global image/text geometry is not itself evidence of semantic loss (Liang et al., 2022). Conversely, RegionCLIP and GLIP learn explicit region–text or phrase–region correspondence (Zhong et al., 2022; Li et al., 2022), a stronger spatial claim than our probes support. We therefore use *cross-modal semantic accessibility* narrowly: P3 and CAPP measure semantic discrimination in the shared embedding space, not spatial grounding or localization.

3 Structured Semantic Fine-Tuning and Capability Measurement

We evaluate each fine-tuning run at multiple checkpoints, separating the supervision intervention from the capability probe: the same training condition can score very differently depending on whether evaluation retains a class-name anchor, removes it, or tests a controlled semantic counterfactual.

Supervision interventions. We compare four conditions (Figure 1). **Name-only (N)** uses class-name-centered contrastive supervision. **Unstructured (U)** encodes the class name and attributes together in one ordinary text stream. **Structured Semantic Routing (SSR; Appendix A)** uses the same pretrained VLM backbone and descriptive supervision, fine-tuned in every condition, but separates NAME and ATTRIBUTE segments with routed causal attention during training. ATTRIBUTE routes have no direct NAME access; for CUB/NAB they may read a shared coarse CONCEPT such as "bird," which is context only and is not fused (ImageNet uses no CONCEPT segment). A training-only Structured Text Fusion combines the NAME summary, a pooled ATTRIBUTE summary, and the EOT summary. **SSR+ND** is identical to SSR except that the NAME-summary contribution is dropped with probability $p{=}0.5$ at training-time fusion, with the remaining fusion weights renormalized.

All methods are evaluated identically. P1, P2, P3, and CAPP use the standard CLIP causal text encoder and the ordinary EOT representation; structured tokens, routed attention, attribute pooling, and Structured Text Fusion are used only during SSR/SSR+ND training. Canonical SSR routing masks are shown in Figure S1; the FULL/PARTIAL/NO name-access masks used in the ablation are shown in Figure S2.

Capability probes. **P1** (class-name retrieval) uses `a photo of a {class_name}` (Recall@1) and measures class-identity accessibility. **P2** (name-and-attribute retrieval) uses `a photo of a {class_name} with {attributes}`, encoded through the ordinary EOT pathway (not the fusion output); because both identity and descriptive information are present, P2 is not attribute-isolating. **P3** (name-free attribute-profile retrieval) removes the class-name anchor and measures class-level attribute accessibility through the ordinary EOT pathway. The exact templates are `a bird with {attributes}` (CUB/NAB), `a butterfly with {attributes}` (FineView), and the frozen canonical ImageNet template `a object with {attributes}`, used exactly as evaluated; P3 does not establish localization, spatial grounding, or part binding. **CAPP** (Counterfactual Attribute-Profile Preference) contrasts the correct full attribute profile with a same-image counterfactual differing in one valid semantic slot/value. Pairwise CAPP uses one alternative; CAPP-Multi ranks the correct value against all valid alternatives for the same semantic slot, with accuracy averaged equally across class–slot pairs. It measures controlled attribute-value discrimination and is complementary to, not interchangeable with, P3. Exact prompts and the CAPP construction are shown in Figure 1c.

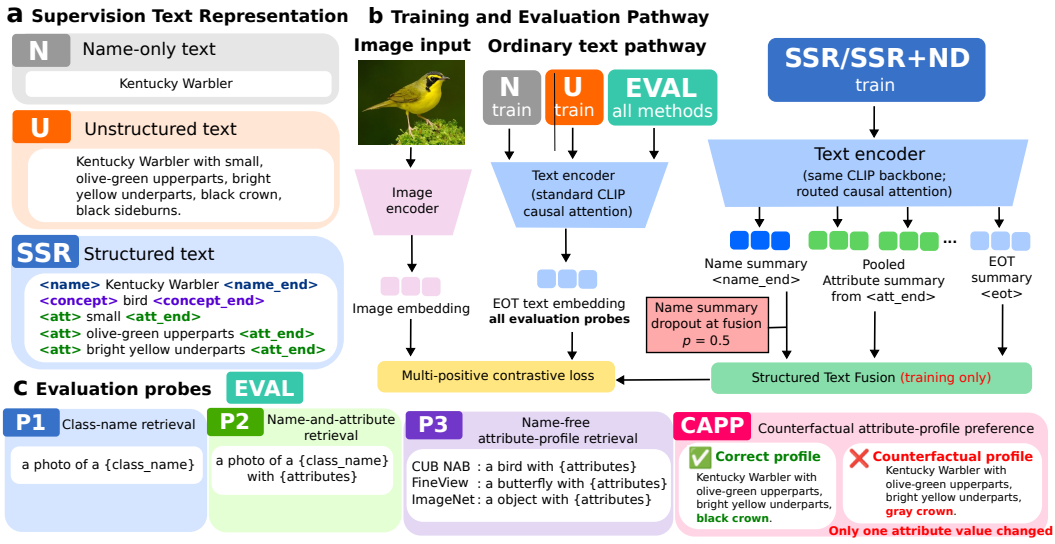


Figure 1: **Supervision representation, training pathway, and capability probes.** **(a)** N uses name-only supervision, U encodes the class name and attributes in one ordinary text stream, and SSR/SSR+ND use structured NAME and ATTRIBUTE segments, with a shared CONCEPT segment for CUB/NAB. **(b)** N/U training and all evaluations use the standard CLIP causal text pathway; SSR/SSR+ND use routed causal attention and a training-only fusion of NAME, pooled ATTRIBUTE, and EOT summaries, with NAME-summary dropout for SSR+ND. **(c)** Exact P1–P3 prompts and the CAPP counterfactual comparison. Photograph: CUB-200-2011.

Acquisition, deployment, and post-peak change. For evaluation distribution d, PT is the pretrained checkpoint evaluated on d; $P3_{\text{peak}}(d)$ is the trajectory maximum and $P3_{\text{deploy}}(d)$ its value at the P1-selected deployment checkpoint:

$$\begin{aligned} \text{Gain}_{\text{peak}} &= P3_{\text{peak}} - P3_{\text{PT}}, \qquad \text{Gain}_{\text{deploy}} = P3_{\text{deploy}} - P3_{\text{PT}}, \\ \text{Forget}_{\text{postpeak}} &= P3_{\text{peak}} - P3_{\text{deploy}}. \end{aligned} \tag{1}$$

The deployment checkpoint is the predefined evaluated checkpoint with maximum *target-domain* P1 for that run; it is never selected using P3, CAPP, or held-out performance, so a semantic optimum may fall before, at, or after the deployment checkpoint. We call post-peak decline a reduction of *acquired* capability only when the trajectory first rose above its pretrained baseline.

Backbones, datasets, and text sources. We evaluate six pretrained dual encoders: DFN5B ViT-H/14-378 and DFN ViT-L/14 (Fang et al., 2024), MetaCLIP ViT-H/14 and ViT-L/14 (Xu et al., 2024), and OpenAI CLIP ViT-L/14 and ViT-B/32 (Radford et al., 2021), on ImageNet (Deng et al., 2009), CUB-200-2011 (Wah et al., 2011), NABirds (Van Horn et al., 2015), FineView Butterfly (Onda & Farrell, 2025), and the ImageNet-V2/-R/-Sketch shifts (Recht et al., 2019; Hendrycks et al., 2021; Wang et al., 2019). CUB/NABirds/FineView descriptions come from public species resources, LLM-processed to normalize length and extract attributes; ImageNet descriptions and attributes are LLM-generated. For reciprocal bird transfer, CUB-trained models are evaluated on the dataset-class-disjoint NAB-348 subset, and NAB-trained models on the dataset-class-disjoint CUB-59 subset; results on the corresponding full populations are reported separately in the supplement. Full protocol details are in Appendix B.

4 Semantic Capability Trajectories

Figure 2a illustrates the trajectory view in a representative CUB-trained model. P1 and P3 do not evolve together: held-out P3 can rise above the pretrained state, peak early, and then decline while target-domain P1 continues toward its own maximum. Thus, the checkpoint selected by class-name retrieval can occur after the best transferable semantic checkpoint. This is clearest for SSR+ND on

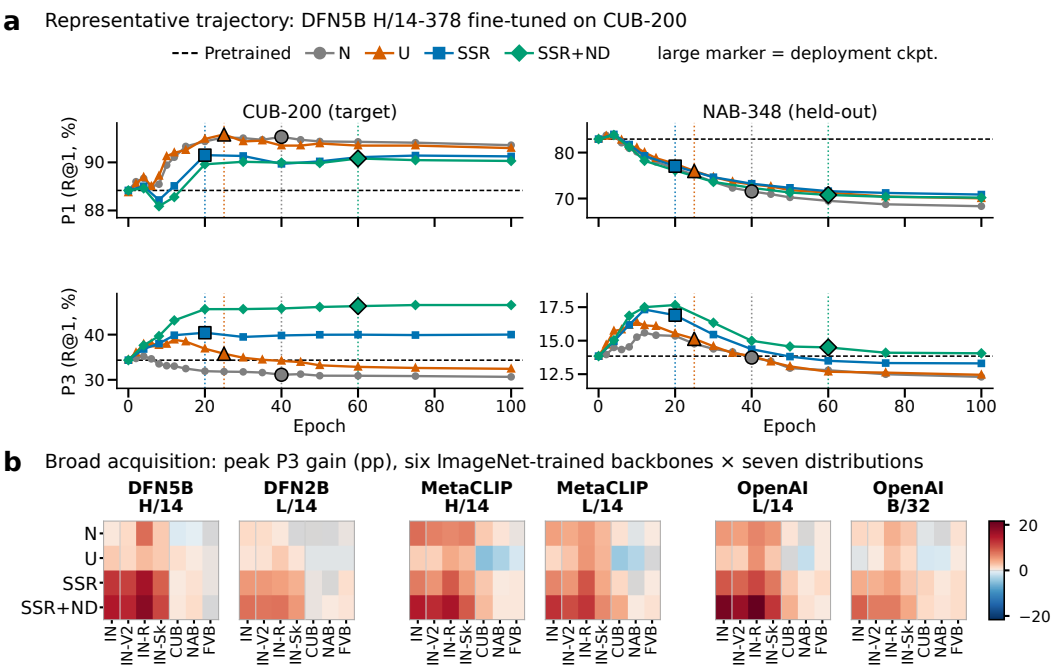


Figure 2: **Capability acquisition and trajectory structure.** **(a)** DFN5B H/14-378 CUB→NAB trajectories; large markers = P1-selected deployment checkpoints. **(b)** Peak P3 gain over pretrained, six ImageNet-trained backbones × seven evaluation distributions.

held-out NAB-348, where held-out P3 peaks around epoch 20 while the P1-selected deployment checkpoint is epoch 60; other methods and domains can peak before, at, or after their deployment checkpoint, so the general point is that capability-specific optima need not coincide with the P1-selected checkpoint. Target and held-out semantic capabilities can also peak at different stages.

4.1 Capability Acquisition

Can fine-tuning acquire semantic capability beyond the pretrained state? We define acquisition as $\mathrm{Gain}_{\mathrm{peak}} = P3_{\mathrm{peak}} - P3_{\mathrm{PT}}$ (§3), a retrospective measure of whether a capability ever emerged, not a deployment rule.

Figure 2b shows that fine-tuning can acquire P3 beyond the pretrained state across six heterogeneous ImageNet-trained backbones, which we treat as distinct models rather than stochastic replicates, and seven evaluation distributions (ImageNet val, three same-ontology ImageNet shifts, and three cross-ontology fine-grained datasets). The largest gains occur on the target and same-ontology ImageNet evaluations, where SSR and especially SSR+ND generally produce much larger gains than N or U; U's mean gain is much smaller and positive in most, but not all, backbones. Cross-ontology gains are substantially smaller and more variable, although structured supervision acquires positive P3 more consistently than N/U overall; the ordering of SSR and SSR+ND is not uniform on every held-out dataset, and no condition is superior in every cell (Table S15).

Held-out acquisition is not restricted to structured supervision: in Figure 2a, even name-only fine-tuning transiently raises held-out P3 above the pretrained baseline (peaks of roughly 15.6–17.7 versus about 13.9), without implying direct attribute supervision. Timing mismatches are analyzed systematically in §4.3.

A weak deployed score therefore does not establish that a capability was never learned; the next question is what determines which capability is acquired.

	PT	N	U	SSR	SSR+ND
P1	79.7	**85.8**	**85.8**	84.9	84.7
P2	74.1	81.1	**86.0**	82.0	82.4
P3	29.9	32.2	28.1	35.6	**44.6**

Table 1: **ImageNet deployment profile.** Six-backbone median R@1 (%) on ImageNet val at each run's P1-selected checkpoint. Per-backbone values: Table S19.

	PT	N	U	SSR	SSR+ND
CUB Pairwise	70.8	70.9	69.9	72.3	**74.7**
CUB CAPP-Multi	45.0	44.6	42.4	46.8	**49.9**
NAB Pairwise	63.4	60.6	63.3	64.7	**65.8**
NAB CAPP-Multi	31.6	28.4	30.2	32.5	**33.9**

Table 2: **Target CAPP summary.** Six-backbone median (%) at each run's P1-selected deployment checkpoint. Per-backbone values: Tables S22–S23.

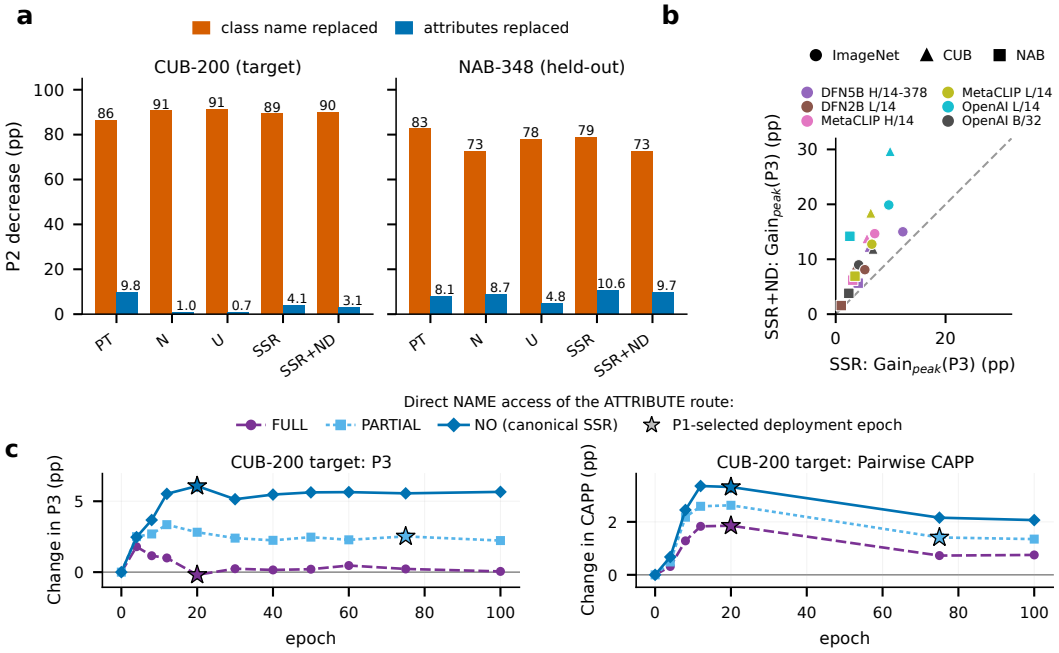


Figure 3: **Why supervision changes what is acquired.** **(a)** P2 decrease when only the class name or only the attributes are replaced; DFN5B H/14-378, CUB-trained (CIs: Table S20). **(b)** SSR+ND vs. SSR peak target P3 gain, 18 backbone×domain pairs (Table S21). **(c)** Direct name-access ablation on the CUB target (DFN5B/CUB; one run per variant). Restricting ATTRIBUTE-route NAME access increases target P3 and Pairwise CAPP; NO is canonical SSR. Stars mark P1-selected deployment; PARTIAL deploys later, so this is a trajectory rather than matched-deployment comparison. Full trajectories: Figure S4.

4.2 Supervision Shapes the Acquired Capability Profile

We next examine whether adding descriptive information is sufficient, or whether its representation and routing determine what is learned. Table 1 shows that the four interventions reach different deployed capability profiles: P1 remains broadly strong across conditions, U reaches the highest P2, and P3 diverges sharply, with SSR and especially SSR+ND reaching much higher values than PT/N/U.

High P2 does not imply strong attribute use. Changing only the class name reduces P2 by 73–91 percentage points across all five conditions on CUB-200 and strict held-out NAB-348, whereas changing only the attributes reduces it by roughly 1–11 points (Figure 3a; DFN5B H/14-378 CUB-trained run; 95% bootstrap CIs over classes, 1,000 resamples, in Table S20). Thus U's high P2 largely reflects class-name reliance, consistent with its comparatively weak deployed P3. SSR and SSR+ND are more attribute-sensitive than U but remain strongly name-dependent, showing increased attribute use alongside, rather than instead of, identity. This is a single-run result, not a six-backbone claim.

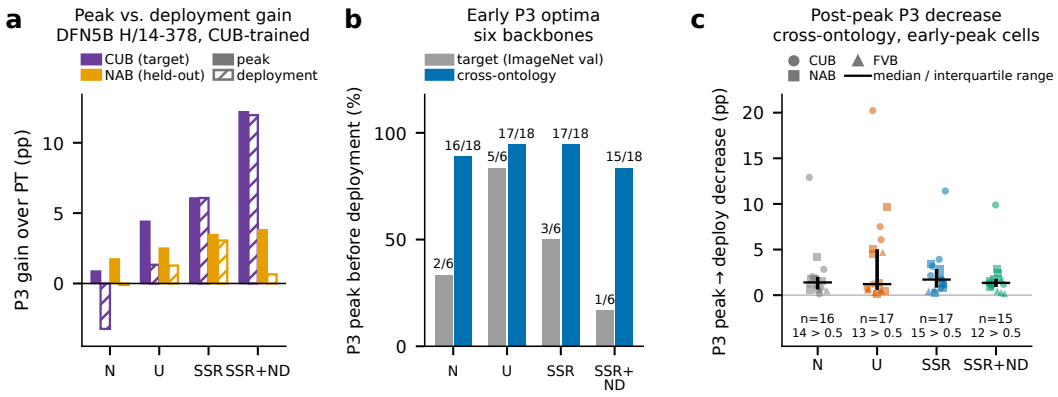


Figure 4: **Capability-specific timing and specialization.** **(a)** Representative peak versus deployment P3 gain for DFN5B H/14-378 fine-tuned on CUB-200. **(b)** Fraction of target (ImageNet val) and cross-ontology evaluations whose P3 optimum precedes deployment, by method across six backbones. **(c)** Peak-to-deployment P3 decrease for cross-ontology trajectories whose P3 optimum precedes deployment; n varies because cells peaking at deployment have no post-peak interval.

How the identity contribution is handled during training – not just the text content – determines how much capability is acquired. SSR and SSR+ND receive identical structured text and use identical routing; SSR+ND additionally drops the NAME channel at fusion with probability 0.5 during training (§3). Across six backbones and three training domains, SSR+ND reaches a larger peak target P3 gain than SSR in all 18 tested backbone×domain pairs (Figure 3b; full values in Table S21). Making the identity contribution less consistently available at training-time fusion changes how much name-free capability is acquired, not only how it is later retained.

CAPP provides complementary evidence by testing controlled attribute discrimination rather than profile retrieval. At each run's own P1-selected deployment checkpoint, SSR+ND has the highest six-backbone median among the fine-tuned conditions in all four target-CAPP rows (Table 2). On the CUB target ontology, SSR improves Pairwise CAPP over U in all six backbones and SSR+ND improves it further in all six, and CAPP-Multi gives the same broad conclusion (Tables S22–S23); the NAB target is less uniform across backbones, and cross-ontology CAPP is more variable still (Table S24). Thus P3 and CAPP support the same broad conclusion while measuring different capabilities; they are not interchangeable.

In a single DFN5B/CUB run, isolating the ATTRIBUTE route from direct class-name access (Figure S2) improves target P3 and target CAPP (Figure 3c); full target/held-out results and additional diagnostics are reported in Tables S25–S28 and Figure S4. However, SSR does not eliminate identity reliance: P2 remains strongly dependent on the class name, as shown by the class-name replacement test.

Together, these results show that adding attribute text and changing how that information is routed are different interventions. U sees descriptive attributes but remains strongly name-dominated in P2. SSR produces stronger name-free P3, while target CAPP improves most consistently on CUB; making the NAME contribution unreliable during training with SSR+ND further increases target P3 acquisition across all 18 backbone×domain comparisons. Which capability is acquired is only part of the trajectory; its own optimum need not coincide with the checkpoint selected by target P1.

4.3 Continued Optimization Can Produce Semantic Specialization

After a semantic capability is acquired, we distinguish acquisition ($\mathrm{Gain}_{\mathrm{peak}}$), what remains at deployment ($\mathrm{Gain}_{\mathrm{deploy}}$), and their difference ($\mathrm{Forget}_{\mathrm{postpeak}} = P3_{\mathrm{peak}} - P3_{\mathrm{deploy}}$, §3).

Figure 4a illustrates why peak and deployment must be separated: a capability can be acquired strongly and still be reduced before the P1-selected checkpoint. The panel shows the representative DFN5B H/14-378 CUB-trained run. N acquires target and held-out P3 above pretrained at its peak, but by deployment its held-out gain has essentially disappeared and target P3 is below the pretrained

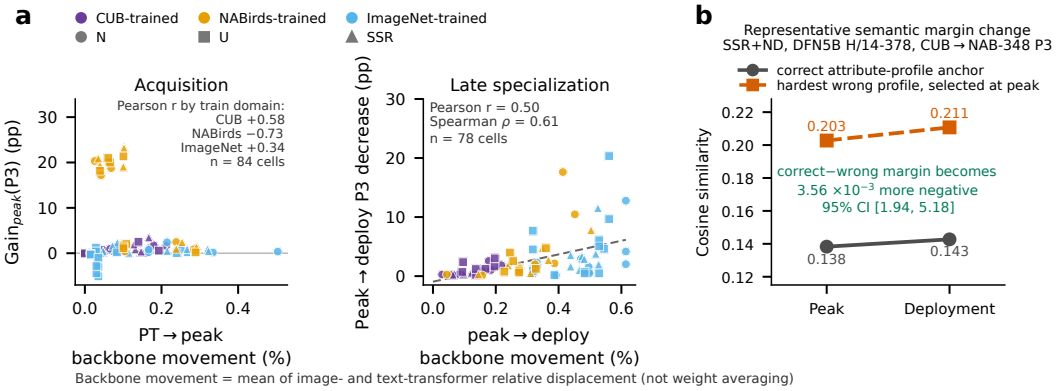


Figure 5: **Late semantic specialization.** **(a)** Joint backbone movement before the P3 peak versus acquisition, and from peak to deployment versus the post-peak P3 decrease, across held-out N/U/SSR trajectories; r = Pearson, ρ = Spearman, n = trajectory cells. **(b)** Representative fixed-competitor analysis for SSR+ND, DFN5B H/14-378, CUB→NAB-348. The peak-selected hardest wrong class is held fixed at deployment; the correct-minus-wrong margin decreases by 0.00356 (95% class-bootstrap CI $[0.00194, 0.00518]$). Because the competitor is the hardest wrong profile selected at the peak, its mean similarity can already exceed that of the correct profile; the analysis tests the subsequent change in their fixed-competitor margin.

level. SSR retains target P3 almost completely, and SSR+ND acquires the largest target P3 gain of any condition; its held-out NAB P3, however, peaks around epoch 20, before the epoch-60 deployment checkpoint, while its target CUB P3 peaks later, around epoch 75. Target and held-out capabilities can therefore specialize differently. In this run, the P3 and Pairwise-CAPP peaks also need not coincide with each other or with deployment (Appendix Figure S10).

Across the six backbones and four methods, cross-ontology P3 peaks before deployment in 65 of 72 cells, compared with 40 of 72 same-ontology shifts and 11 of 24 target-domain cells (Figure 4b plots the target and cross-ontology fractions; same-ontology counts are in Appendix F). This pattern is consistent across methods, with 15–17 of 18 cross-ontology cells peaking early for each method. Among these 65 early-peaking trajectories, 54 lose more than 0.5 percentage points of P3 by deployment; the 0.5-point cutoff is a descriptive robustness threshold, not a significance test. The median decline is similar across methods, ranging from 1.21 to 1.71 points (Figure 4c). Thus, transferable P3 often reaches its best value before the target-P1-selected deployment checkpoint and then declines with continued optimization, while target-domain P3 more often peaks at the deployment checkpoint. Full timing results are provided in Appendix F.

4.4 Understanding Late Specialization

To characterize the transition from a capability's peak to a later specialized state, we first test whether parameter movement tracks acquisition and the post-peak decrease across the broader trajectory grid, then examine the geometry of the held-out decrease in a representative transition.

Broad association. We next ask whether parameter movement accompanies acquisition or later decline. In a separate movement-analysis grid covering four backbones and N/U/SSR (Appendix G.1; SSR+ND is not included), we measure movement as the arithmetic mean of matched relative displacement in the pretrained image and text transformers. Pretrained-to-peak movement has no consistent relationship with P3 acquisition: the association changes sign across training domains (Figure 5a, left). In contrast, among held-out trajectories that peak before deployment, greater peak-to-deployment movement is associated with a larger P3 decrease ($r = 0.50$, $\rho = 0.61$, $n = 78$; Figure 5a, right). Image- and text-encoder movement are strongly correlated ($r = 0.85$), and which encoder appears more strongly associated with the post-peak decrease depends on how movement is scaled across backbones; we therefore do not attribute the effect to either encoder individually. These are descriptive associations across heterogeneous backbones, not causal estimates or stochastic replicates. Accordingly, this association should not be interpreted as evidence that the same

movement relationship holds for SSR+ND. The relationship is not specific to P3: over the same intervals, movement is even more strongly associated with held-out P1 decline ($r = 0.73$, $\rho = 0.79$; Appendix G.2).

Representative geometry. Figure 5b examines the representative SSR+ND CUB→NAB-348 transition. From the P3 peak to deployment, the peak-selected wrong attribute profile moves closer to the image more than the correct profile does, reducing the correct-versus-wrong margin on the held-out data (Appendix G.8). Supplementary analyses support and bound this interpretation. The effect extends across the wrong-class similarity spectrum rather than being specific to the hardest competitor (Appendix G.3), but the exact signature is not universal: an independent text-frozen U control does not exhibit the same resolved geometric signature (Appendices G.8–G.9). Image-only class structure remains substantially preserved across pretrained, peak, and deployment states (Appendix G.10), and coarse geometry hypotheses do not fully explain the effect (Appendix G.11). Together, these results suggest that the representative late-stage decline reflects reduced cross-modal semantic accessibility rather than broad erasure of visual class information, while also showing that late-stage semantic degradation does not always arise through the same embedding-geometry change.

5 Discussion and Limitations

Viewing fine-tuning as a capability trajectory rather than a single pretrain-versus-deployment comparison changes what a weak deployed score means: the capability may never have been acquired, or it may have emerged earlier and later become less accessible. Our results show both cases. Unstructured descriptive supervision can acquire little transferable name-free capability, whereas structured supervision can acquire substantial capability whose optimum precedes deployment.

Two additional controls help bound the late-specialization interpretation (Appendix H). In the representative DFN5B/CUB setting, WiSE interpolation partially recovers held-out P3 for N/U while preserving strong target P1/P3, showing that the decline need not be irreversible. We do not apply this interpolation to SSR/SSR+ND because their structured components lack direct pretrained counterparts. Separately, a target-side hard-negative objective improves target P3 without improving held-out P3 at matched later checkpoints, showing that stronger in-domain semantic discrimination does not necessarily translate into stronger transferable capability.

Limitations. The §4.4 geometric diagnostics are representative single-run analyses, and the broader movement analysis is descriptive rather than causal; neither supports encoder-specific attribution. Main trajectory cells are also single runs, so variation across the six backbones reflects model heterogeneity rather than training stochasticity. Our comparisons are controlled supervision/routing interventions rather than a benchmark against task-specific fine-grained methods. Deployment is selected by target P1, and P2/P3/CAPP have the scope limitations defined in §3. Dataset-text sources differ across domains, so residual source, template, or generation effects cannot be excluded. Additional limitations are detailed in Appendix J.1.

6 Conclusion

We studied VLM fine-tuning as a semantic capability trajectory rather than a single pretrain-to-deployment endpoint. Across six pretrained backbones, fine-tuning can acquire capabilities beyond the pretrained state, and Structured Semantic Routing, particularly with NAME-summary dropout, strengthens name-free attribute-profile capability in most tested settings, showing that how descriptive supervision is represented and routed changes what is acquired. Different capabilities reach their optima at different stages: transferable name-free attribute capability often peaks before the checkpoint selected by target class-name retrieval. In a representative diagnostic study, the subsequent decline was consistent with reduced cross-modal semantic accessibility while substantial image-only class structure remained. Tracking capability trajectories therefore distinguishes failure to acquire from post-acquisition decline.

AI USE STATEMENT

In this work, we used generative AI tools for tasks requiring disclosure, including providing feedback on research methodology and experimental design, assisting with software and analysis code, assisting in the interpretation and organization of experimental results, and processing dataset text annotations. For CUB/NABirds and FineView, source descriptions were collected from public web resources and generative AI was used to normalize/shorten the descriptions and extract individual observable visual attributes. For ImageNet, class-level descriptions and visual attributes were generated with generative AI.

Additionally, we used generative AI tools to summarize and synthesize author-provided literature, brainstorm and organize the manuscript, suggest manuscript structure, draft and edit manuscript text for clarity, assist with scientific figure planning, verify and format bibliographic information, and propose titles and keywords.

The literature corpus used for the paper was selected and provided by the authors rather than discovered by generative AI. We reviewed all AI-assisted work. Quantitative results were obtained from our experimental runs and checked against the corresponding experiment outputs and provenance records. AI-processed dataset annotations were reviewed as part of the dataset construction and experimental pipeline. Literature-based claims and citations were checked against the source papers, and AI-assisted code and analyses were reviewed and validated by the authors. We take responsibility for the final content of this work, including text, claims, dataset annotations, code, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 defines the supervision conditions, semantic capability probes, checkpoint-selection protocol, and trajectory quantities used throughout the paper. The supplementary material provides additional training and evaluation details, exact probe definitions, checkpoint schedules, per-backbone results, evaluation populations, and uncertainty analyses where applicable. We report the evaluated checkpoints and pretrained baselines separately for each evaluation distribution so that the acquisition, deployment, and post-peak quantities can be reconstructed from the underlying results.

REFERENCES

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

Md Atabuzzaman, Andrew Zhang, and Chris Thomas. Zero-shot fine-grained image classification using large vision-language models. *arXiv preprint arXiv:2510.03903*, 2025.

Jessica Bader, Leander Girrbach, Stephan Alaniz, and Zeynep Akata. Sub: Benchmarking cbm generalization via synthetic attribute substitutions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23188–23198. IEEE, 2025.

Ethan Baron, Idan Tankel, Peter Tu, and Guy Ben-Yosef. Real classification by description: Extending clip's limits of part attributes recognition. *arXiv preprint arXiv:2412.13947*, 2024.

Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22520–22529. IEEE, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, pp. 36221–36237, 2024.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19338–19347. IEEE, 2023.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 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. IEEE, 2021.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M Asano, Nanne Van Noord, Marcel Worring, and Cees GM Snoek. Tulip: Token-length upgraded clip. *arXiv preprint arXiv:2410.10034*, 2024.

Suguru Onda and Ryan Farrell. The fineview dataset: A 3d scanned multi-view object dataset of fine-grained category instances. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5623–5634. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Karsten Roth, Jae Myung Kim, A Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15700–15711. IEEE, 2023.

Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjhieva, and Yue Dong. Vlms need words: Vision language models ignore visual detail in favor of semantic anchors. *arXiv preprint arXiv:2604.02486*, 2026.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961. IEEE, 2022.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. Fg-clip 2: A bilingual fine-grained vision-language alignment model. *arXiv preprint arXiv:2510.10921*, 2025a.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025b.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference on Learning Representations*, volume 2024, pp. 47812–47831, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112. IEEE, 2022.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782. IEEE, 2022.

A SSR Architecture and Objective

This appendix provides detailed method, protocol, and numerical documentation for the paper. Sections A–C describe the method, data, probes, and training recipes; Sections D–H give the full numerical support for every main-paper figure and claim; Section J maps each main-paper item to its supplement table, population, checkpoint rule, and source artifact. Six-backbone results are descriptive across heterogeneous pretrained models and are never treated as stochastic replicates. Sections marked "single run" report one training run. The formulation below is consistent with the current implementation and the archived parameter records for the reported runs.

A.1 Structured semantic representation and routed attention

Let $\mathcal{D} = \{(I_i, y_i)\}$ be a training set. Each class c has a class name n_c and attributes $\mathcal{A}_c = \{a_{c,1}, \ldots, a_{c,K_c}\}$. For the CUB/NAB canonical SSR, with shared coarse concept q_c (e.g., "bird"), the structured sequence is

$$x_c = \big[\texttt{<sot>},\ \texttt{<name>}\, n_c\, \texttt{<name_end>},\ \texttt{<concept>}\, q_c\, \texttt{<concept_end>},\ \{\texttt{<att>}\, a_{c,k}\, \texttt{<att_end>}\}_{k=1}^{K_c},\ \texttt{<eot>}\big],$$

for example `<name> Kentucky Warbler <name_end> <concept> bird <concept_end> <att> small <att_end> <att> olive-green upperparts <att_end> ...`. The concept is a coarse semantic context shared across the fine-grained classes of the dataset; it is not the class identity. For ImageNet canonical SSR there is no concept segment:

$$x_c = [\texttt{<sot>},\ \texttt{<name>}\, n_c\, \texttt{<name_end>},\ \{\texttt{<att>}\, a_{c,k}\, \texttt{<att_end>}\}_{k=1}^{K_c},\ \texttt{<eot>}].$$

ImageNet training uses no concept token. The word "object" occurs only in the P3 *evaluation* template ("a object with {attributes}") and never as a training-time concept.

Routed attention. Every position attends to <sot> and itself, and attention inside a segment is causal. Let r^{name}, r_k^{att}, and r^{eot} denote the positions of <name_end>, the k-th <att_end>, and <eot>.

- **CUB/NAB (shared-concept routing).** The NAME route reads only the NAME segment. CONCEPT tokens read only the CONCEPT segment. The k-th ATTRIBUTE route reads its own segment *and the shared CONCEPT content*, but not the NAME segment and not other attribute segments. <eot> reads only the summary endpoints $\{r^{\text{name}}\} \cup \{r_k^{\text{att}}\}_k$.
- **ImageNet (no-concept routing).** As above without a CONCEPT segment: the k-th ATTRIBUTE route reads only its own segment, there is no NAME-to-ATTRIBUTE access, and <eot> reads only $\{r^{\text{name}}\} \cup \{r_k^{\text{att}}\}_k$.

<eot> does not attend to every raw token and does not read the concept boundary token. The exact masks are given in §A.4.

Relation to earlier SSR routing variants. Earlier SSR routing variants allowed each attribute route to access the class-name segment directly and used no concept segment. In the current canonical CUB/NAB formulation, we instead provide the attribute routes with a shared coarse concept (e.g., "bird") while removing their direct access to the fine-grained class name. This provides coarse semantic context without giving the attribute route direct access to the fine-grained class identity; it does not imply complete disentanglement, and class-name information still reaches the fused representation through the NAME summary (and ordinary text-encoder P2 remains strongly name-dominated; §E). ImageNet uses the corresponding no-concept routing. Some historical runs (§C.6) used the earlier routing and are never pooled with the canonical one.

A.2 Attribute pooling, structured fusion, and name-branch dropout

Passing x_c through the routed text encoder gives final-layer hidden states H_c, from which three kinds of features are read at the boundary tokens:

$$h_c^{\text{name}} = H_c[\texttt{<name_end>}], \qquad h_{c,k}^{\text{att}} = H_c[\texttt{<att_end>}_k], \qquad h_c^{\text{eot}} = H_c[\texttt{<eot>}].$$

No concept feature is extracted. The concept affects the attribute features only through the routed attention.

Attribute Importance Pooling (AIP). The ATT_END features are pooled into one attribute summary with a learned importance score (GELU σ; padded attributes masked before the softmax):

$$\beta_{c,k} = \frac{\exp\big(w_2^\top \sigma(W_1 h_{c,k}^{\text{att}})\big)}{\sum_{j=1}^{K_c} \exp\big(w_2^\top \sigma(W_1 h_{c,j}^{\text{att}})\big)}, \qquad h_c^{\text{attr}} = \sum_{k=1}^{K_c} \beta_{c,k} h_{c,k}^{\text{att}}.$$

AIP is applied to the final-layer hidden states before the text projection.

StructuredTextFusion (training only). The three channels (NAME, pooled ATTRIBUTE, EOT) are passed through the text projection P and normalized, $\tilde{h}_c^{\text{ch}} = \text{norm}(P\, h_c^{\text{ch}})$ for $\text{ch} \in \{\text{eot}, \text{name}, \text{attr}\}$. A gate over the concatenated channels predicts fusion weights, and the structured training text embedding is the normalized weighted sum:

$$\alpha_c = \text{softmax}\Big(W_2\, \sigma\big(W_1[\tilde{h}_c^{\text{eot}}; \tilde{h}_c^{\text{name}}; \tilde{h}_c^{\text{attr}}]\big)\Big) \in \mathbb{R}^3, \qquad t_c = \text{norm}\Big(\alpha_c^{\text{eot}} \tilde{h}_c^{\text{eot}} + \alpha_c^{\text{name}} \tilde{h}_c^{\text{name}} + \alpha_c^{\text{attr}} \tilde{h}_c^{\text{attr}}\Big).$$

The fusion inputs are only the NAME, ATTRIBUTE, and EOT summaries; the concept is *not* a fusion input. The learned gate is a two-layer network ($3D \to D \to 3$).

SSR+ND. SSR+ND uses the same structured sequence, routed encoder, and routing mask as SSR. During training, for each sample independently with probability $p = 0.5$, the NAME channel's gate logit is set to $-\infty$ before the softmax, so $\alpha_c^{\text{name}} = 0$ and the EOT and ATTRIBUTE weights are renormalized by the softmax. The NAME tokens remain in the encoded sequence, the routing mask is unchanged, and the NAME feature still enters the gate network's input; only its fusion weight is removed. Dropout is applied in training only.

Inference and evaluation. At inference and for every capability probe (P1, P2, P3, CAPP), plain prompts are encoded by the ordinary causal text encoder and the feature at the EOT position is projected (the "ordinary EOT text embedding"); no structured tokens, routing mask, AIP, or fusion are used. The structured fusion is a training-time construct only, and the earlier formulation's fused-representation inference and attribute-patch reranking are not part of the evaluation reported here. N and U use the ordinary text pathway in both training and evaluation.

A.3 Multi-positive CLIP-style contrastive objective

For a training batch of image embeddings $Z \in \mathbb{R}^{B\times d}$ with labels y_i and the corresponding (structured, for SSR/SSR+ND) text embeddings $T \in \mathbb{R}^{B\times d}$, let $S = \gamma Z T^\top$ with the learnable CLIP logit scale γ. Because several images in a batch share a class, the one-hot CLIP target is replaced by the multi-positive target

$$Q_{ij}^{I\to T} = \frac{\mathbf{1}[y_i = y_j]}{\sum_{k=1}^{B} \mathbf{1}[y_i = y_k]}, \qquad \mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} Q_{ij}^{I\to T} \log \frac{\exp(S_{ij})}{\sum_{k=1}^{B}\exp(S_{ik})},$$

with $\mathcal{L}_{T\to I}$ defined symmetrically, and

$$\mathcal{L}_{\mathrm{mp}} = \tfrac{1}{2}\big(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\big).$$

This is a multi-positive objective, not single-positive InfoNCE. In the implementation the targets are tiled $1/(\text{positives per class})$ entries over class-balanced effective batches (200 classes $\times$ 10 images for CUB, 555 $\times$ 4 for NAB), which equals the definition above; the contrastive loss spans the full effective batch across gradient-accumulation steps. N, U, SSR, and SSR+ND all use this same loss. The remaining recipe settings are in §C.2.

A.4 Canonical semantic routing masks

Routing sets. Let S^{name}, S^{con}, and S_k^{att} be the token positions of the NAME, CONCEPT, and k-th ATTRIBUTE segments. Every position attends to <sot> and to itself; within a segment attention is causal. The additional access of each route is:

- **Shared-concept routing (CUB/NAB).** NAME tokens (including r^{name}) read only earlier NAME tokens. CONCEPT tokens read only earlier CONCEPT tokens. ATTRIBUTE$_k$ tokens (including r_k^{att}) read earlier tokens of their own segment *and the CONCEPT content tokens* (excluding the concept boundary token), but neither the NAME segment nor other attribute segments. <eot> reads only $\{r^{\mathrm{name}}\} \cup \{r_k^{\mathrm{att}}\}_k$; it does not read the concept boundary token or content tokens directly.
- **No-concept routing (ImageNet).** As above with no CONCEPT segment: ATTRIBUTE$_k$ tokens read only their own segment (the attribute-to-name link is off), and <eot> reads only $\{r^{\mathrm{name}}\} \cup \{r_k^{\mathrm{att}}\}_k$.

Consequently the ATTRIBUTE routes have no direct access to the class name, while name information reaches the fused representation through r^{name} (via <eot> and the NAME channel).

The concept segment. In the shared-concept routing, a single coarse category word (e.g., "bird" for CUB and NAB) is placed between the NAME and ATTRIBUTE segments. Every ATTRIBUTE route reads it, so that each attribute phrase ("black crown") is interpreted in the same coarse category context without exposing the class name; the concept is shared across all classes of the dataset and carries no class identity. It is context for the attribute routes and is not a fusion input; <eot> does not read it directly, so it influences the class embedding only through the attribute summaries. ImageNet has no single coarse category, so its routing has no concept segment.

Figure S1 shows the token-level access pattern. SSR and SSR+ND use the identical mask; SSR+ND's name dropout acts at the fusion input, not in the mask. An implementation summary of these routings is in Table S6.

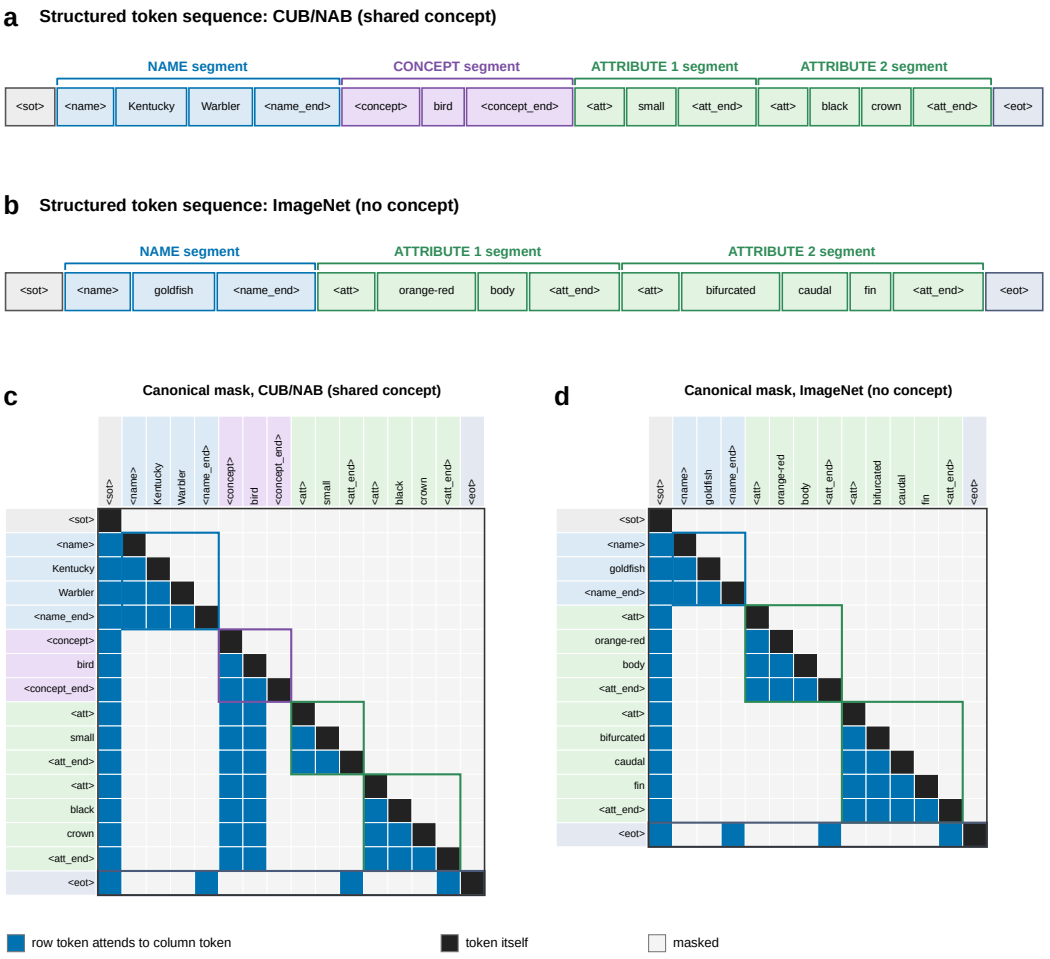


Figure S1: Structured token sequences and canonical routed-attention masks. **(a)** CUB/NAB sequence with the shared concept, showing every start/end boundary token (<sot>, <name>/<name_end>, <concept>/<concept_end>, <att>/<att_end>, <eot>); the example attributes are "small" and "black crown". **(b)** ImageNet sequence without a concept (class "goldfish", attributes "orange-red body" and "bifurcated caudal fin"). **(c, d)** Token-level masks built from the mask definitions used in training: a blue cell means the row token attends to the column token, black is the token itself, light grey is masked. Every token attends <sot>; attention within a segment is causal. In (c), ATTRIBUTE tokens also read the CONCEPT tokens (<concept> and its content, but not <concept_end>) and never the NAME segment or other attribute segments; <eot> reads only <name_end> and the <att_end> tokens. In (d) there is no concept and ATTRIBUTE tokens read only their own segment. "object" is a P3 evaluation-template word only, never a training token.

A.5 Controlled name-access variants

Figure S2 shows the three variants. They are controlled leakage/name-access analysis variants, not standard SSR recipes, and are used only for the analysis in §E.7 (Tables S25–S28). In all cases EOT remains name-accessible through NAME_END.

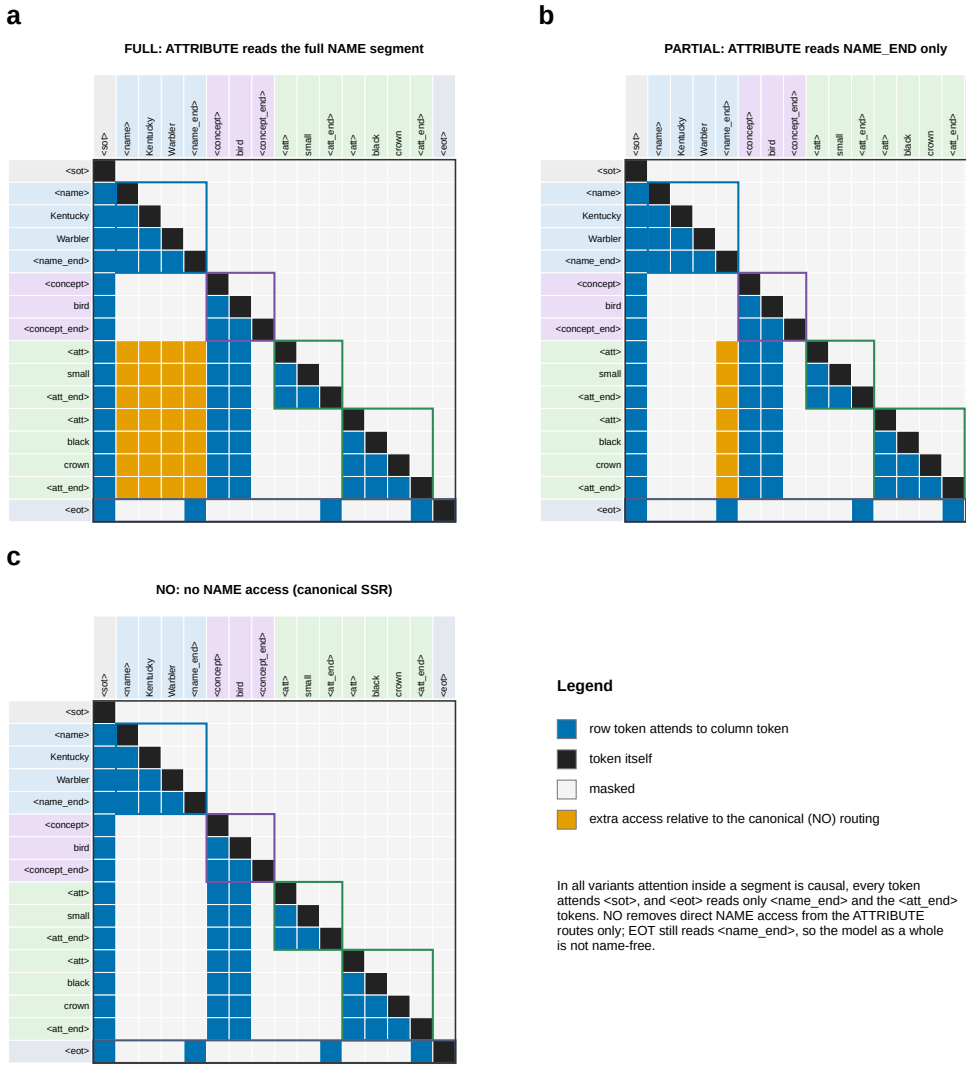


Figure S2: Token-level masks for the controlled name-access variants (direct name-access ablation; CUB/NAB shared-concept setup; same conventions as Figure S1). Orange cells mark access beyond the canonical routing. **(a) FULL:** ATTRIBUTE tokens also read all NAME tokens (<name>, content, <name_end>). **(b) PARTIAL:** ATTRIBUTE tokens also read only the <name_end> boundary token, not the class-name content. **(c) NO:** no NAME access; identical to the canonical SSR routing. In all variants <eot> still reads <name_end>, so NO removes direct NAME access from the ATTRIBUTE route only and the model as a whole is not name-free.

B Datasets, Text Construction, and Evaluation Protocol

B.1 Datasets and source descriptions

We use ImageNet-1k (Deng et al., 2009) and its same-ontology shifts ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021), and ImageNet-Sketch (Wang et al., 2019); CUB-200-2011 (Wah et al., 2011) (200 classes); NABirds (Van Horn et al., 2015) (555 classes as used here); and FineView Butterfly (Onda & Farrell, 2025) (130 classes). ImageNet-trained models are evaluated on ImageNet val (target), the three shifts (same ontology), and CUB-200, NAB-348, and

Dataset / class	Raw source (excerpt)	Extracted attributes used	P1 / P3 strings
CUB: Kentucky Warbler	All About Birds: "Adult male has olive green upperparts, bright yellow underparts. Forehead, crown, and side of head black . . . "	small; olive-green upperparts; bright yellow underparts; black crown; black sideburns; yellow spectacles; no wing bars	P1: "a photo of a Kentucky Warbler"; P3: "a bird with {attributes}"
NAB: female or juvenile Lesser Goldfinch	All About Birds: "tiny, stub-billed songbirds with long, pointed wings, and short, notched tails . . . olive backs, dull yellow underparts, and black wings marked by two whitish wingbars"	stub bill; long pointed wings; short notched tail; olive back; dull yellow underparts; black wings; whitish wingbars	P3: "a bird with {attributes}"
FineView: Monarch	BAMONA: "Upperside of male is bright orange with wide black borders and black veins . . . white spots on borders and apex"	orange upperside; wide black borders; black veins; blurred black veins; white border spots; white wing-tip spots	P3: "a butterfly with {attributes}"
ImageNet: Goldfish	(LLM-generated; no raw source)	orange-red body; bright yellow body; torpedo-shaped body; large reflective scales; bifurcated caudal fin; stiff dorsal fin; smooth operculum	P3: "a object with {attributes}"

Table S1: Verified examples of source description, extracted attributes, and probe strings. NAB classes are plumage forms (e.g., female or juvenile), split from the per-species source text by the LLM. CUB attribute list shown is the CUB-59 strict-set list.

FineView Butterfly (cross-ontology). CUB- and NAB-trained models are evaluated on their own test sets (target) and reciprocally on the other bird dataset (§B.5).

B.2 LLM PROCESSING AND ATTRIBUTE EXTRACTION

Sources. For CUB-200-2011 and NABirds we obtain field-guide-style species descriptions from Cornell Lab's *All About Birds*. For FineView Butterfly we use species descriptions from Butterflies and Moths of North America (BAMONA). Because these descriptions vary in length, style, and visual detail, an LLM was used to normalize or shorten them and to extract individual observable visual attributes; these processed descriptions and attributes form the text inputs of the unstructured and structured conditions and of the semantic probes. For ImageNet, class-level descriptions and visual attributes are generated with an LLM (there is no raw web source). Where applicable, U and SSR/SSR+ND derive their semantic information from the same underlying descriptions.

Verified examples. Table S1 shows one class per dataset with the raw source text (where one exists), the extracted attributes actually used, and the resulting probe strings.

B.3 EXACT CAPABILITY-PROBE DEFINITIONS

All probes retrieve text prompts from images (image-to-text Recall@1 unless stated) through the *ordinary CLIP/EOT text pathway*; none uses structured fusion.

- **P1** (class-name identity): `a photo of a {class_name}`.
- **P2** (joint name+attribute compatibility): `a photo of a {class_name} with {attributes}`. P2 is class-name dominated and is not an attribute-isolating metric.
- **P3** (name-free class-level attribute-profile retrieval): CUB/NAB `a bird with {attributes}`; FineView `a butterfly with {attributes}`; ImageNet family `a object with {attributes}`. The ungrammatical "a object" is the exact string used. P3 is not localization, spatial grounding, or part binding.

The name-swap diagnostic replaces only the class name in the P2 prompt with another class's name; the attribute-swap diagnostic replaces only the attributes with another class's attributes.

B.4 CAPP CONSTRUCTION AND AGGREGATION

CAPP (Counterfactual Attribute-Profile Preference) presents the *same image* with a full correct attribute profile (rendered, e.g., "a bird with black crown, bright yellow underparts, olive-green upperparts") and a full counterfactual profile that differs in exactly one valid semantic slot/value; all other profile content is identical. The counterfactual value is the value of that slot for another class (donor). Pairwise CAPP uses one counterfactual; CAPP-Multi ranks the correct value against all valid same-slot alternatives (top-1 reported). Accuracies are macro-averaged over (class, slot) cells (class-slot-macro). A profile/slot is valid only if the slot is well-defined for the class under the strict-contrast rules used to construct the author-reviewed full-class CAPP profiles. CAPP measures controlled attribute-value discrimination; it is not spatial grounding or localization and is complementary to, not interchangeable with, P3.

```
for each class c, slot s with valid value v_c[s]:
   correct = render(profile(c))
   for each donor value w != v_c[s] (pairwise:  one w; multi:
all valid w):
     counter = render(profile(c) with slot s := w)
     score = [sim(img, correct) > sim(img, counter)] #
multi:  rank of correct
   cell(c,s) = mean over images and donors
CAPP = mean over cells (c,s)
```

B.5 EVALUATION POPULATIONS

CUB-trained models are evaluated cross-ontology on NAB-555 (full) and NAB-348 (strict, dataset-class-disjoint from CUB). NAB-trained models are evaluated on CUB-200 (full) and CUB-59 (strict, CUB classes that do not overlap NABirds under the frozen dataset-class mapping). Full and strict populations are reported separately and never pooled; CUB-to-NAB and NAB-to-CUB are never pooled.

Provenance of the strict subsets. We use frozen strict-overlap subsets containing 348 NABirds classes and 59 CUB classes. The archived class lists and downstream evaluation artifacts are internally consistent, but the original scripts used to derive these subsets were not preserved; we therefore treat the released class lists as the authoritative evaluation artifacts.

Taxonomy note. Because CUB and NABirds use historical dataset taxonomies, some legacy dataset classes may correspond to taxa that have subsequently been renamed, split, or merged in external taxonomic resources. We preserve the original dataset class identities for evaluation; attribute descriptions may therefore be shared when the external source treats those historical classes jointly. The strict split is defined operationally by the original dataset identities and the frozen mapping used in the experiments, and is not claimed to be disjoint under current taxonomy.

B.6 CHECKPOINT AND DEPLOYMENT CONVENTIONS

Deployment is each run's own checkpoint with the highest target-domain P1; it is never selected using P3, CAPP, or held-out performance. CUB/NAB trajectory locations are epochs. ImageNet checkpoints are 11 cumulative-data milestones (Table S4), not epochs and not proportional to gradient updates; the deployment milestone of every ImageNet-trained run, for all six backbones, is the milestone with the highest ImageNet-val P1 measured on the full 50,000-image validation set during training, which is the final milestone (11) for all 24 runs. For OpenAI N/U/SSR the training records of the earliest milestones (1–5 for L/14, 1–6 for B/32) are not archived; these milestones are excluded from the selection, which does not affect it because their P1 on the smaller 5,000-image probe subset is at least 4 points below that of the final milestone. Per-milestone P1 used in the trajectory tables comes from that probe subset, whose argmax differs from the full-validation choice by one milestone for three runs (MetaCLIP H/14 SSR+ND, MetaCLIP L/14 U, OpenAI B/32 SSR; probe margins at most 0.3 pp), so the full-validation rule is used for deployment selection throughout. PT is the pretrained checkpoint evaluated directly on the dataset named in each table block. $\mathrm{Gain}_{\mathrm{peak}} = P3_{\mathrm{peak}} - P3_{\mathrm{PT}}$, $\mathrm{Gain}_{\mathrm{deploy}} = P3_{\mathrm{deploy}} - P3_{\mathrm{PT}}$,

$\text{Forget}_{\text{postpeak}} = P3_{\text{peak}} - P3_{\text{deploy}}$; a post-peak decline is called reduction of an acquired capability only when the trajectory first rose above PT.

B.7 Bootstrap, duplicate-evaluation, and uncertainty conventions

- **P2 name-/attribute-swap intervals (Table S20):** class-level bootstrap, 1000 replicates (classes resampled with replacement, all images of sampled classes kept), 95% percentile intervals, seed 20260917.
- **CAPP paired comparisons and name-access-ablation CAPP contrasts:** paired class-level bootstrap with 10,000 resamples, same classes for both methods. No CAPP intervals appear in the main tables.
- **Geometry correlations (Table S47):** class-level Spearman correlations with 10,000 class resamples for 95% intervals.
- **ImageNet duplicate-evaluation variability:** repeated evaluation of the same checkpoint differs by about 0.18 pp per run (repeated-evaluation checks); differences below roughly 0.5 pp are interpreted cautiously.
- **No inferential tests across backbones or methods are reported.** The six backbones are heterogeneous, not replicates; bootstrap intervals quantify evaluation-set uncertainty, and multi-seed runs (§C.6) quantify training stochasticity only for the configurations actually repeated.

B.8 Zero-shot motivation: description dilution

This subsection documents the pretrained, zero-shot observation cited in the Introduction. It is a motivational analysis of frozen pretrained encoders and is not part of the fine-tuning trajectory experiments; it should not be extended to fine-tuned models. Table S2 reports zero-shot CUB accuracy for two pretrained backbones under three text inputs: the class name alone, a concept-plus-class-name prompt, and the concept-plus-class-name prompt followed by a field-guide-style description of about 77 tokens. For both backbones, accuracy is highest with the class name alone and declines as more text is added. Descriptions longer than the standard 77-token context window can only be processed without truncation by a long-context variant of each backbone (TULIP (Najdenkoska et al., 2024)); both variants decline further as description length grows, more steeply for the DFN-5B-based variant. Providing more semantic text does not by itself make that information more accessible through the resulting representation. The descriptions are unprocessed field-guide text rather than the extracted attribute lists used for P2/P3, and this result does not compare against shuffled or wrong-class text.

Text input	DFN5B ViT-H/14-378	OpenAI ViT-L/14
Class name only	**88.80**	**57.18**
Concept + class name	87.88	55.93
Concept + class name + description (∼77 tokens)	84.77	50.07
Long-context TULIP variant of the same backbone		
Concept + class name + description (200–300 tokens)	36.02	41.50
Concept + class name + description (100–500 tokens)	28.55	39.30

Table S2: Zero-shot CUB retrieval accuracy (%) for two pretrained backbones as text is added to the class-name prompt (frozen encoders, evaluation only). The lower block uses the long-context TULIP variant (Najdenkoska et al., 2024) of each backbone, which can process descriptions beyond the 77-token window without truncation.

Backbone	open_clip model	Pretrained tag
DFN5B H/14-378	ViT-H-14-378-quickgelu	dfn5b
DFN2B L/14	ViT-L-14	dfn2b_s39b
MetaCLIP H/14	ViT-H-14-quickgelu	metaclip_fullcc
MetaCLIP L/14	ViT-L-14-quickgelu	metaclip_fullcc
OpenAI L/14	ViT-L-14-quickgelu	openai
OpenAI B/32	ViT-B-32-quickgelu	openai

Table S3: Pretrained initializations (open_clip model names and pretrained tags read from the run parameter files). Backbone descriptions and parameter counts are in Table S17.

C Training and Implementation Details

C.1 Backbone initialization and pretrained sources

C.2 Training recipes by dataset/domain

All runs use the open_clip trainer with a multi-positive contrastive loss (temperature 1, mixing coefficient 0.5), bf16 mixed precision, per-GPU batch 20 with gradient accumulation, seed 0, no additional augmentation, text context extended from 77 to 600 tokens (dynamic NTK, scale factor 8), and no frozen image encoder. Tables S5 and S4 give the settings read from the run parameter files; they are parent-specific and *not* normalized across runs.

Method	LR	LR_{sl}	WD	Eff. batch	Routing
N	5e-06	5e-06	0.1	2000	–
U	5e-06	5e-06	0.1	2000	–
SSR	5e-06	5e-06	0.1	2000	no-concept
SSR+ND	5e-06	5e-06	0.1	2000	no-concept, p=0.5

Table S4: ImageNet-1k training settings used across the six backbones; available run parameter records agree with these settings, with early OpenAI records unavailable as noted below. Training is a chain of 11 resumed steps, each one epoch over the incremental data slice up to cumulative fractions 0.25, 0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100% (4k to 1.3M images; 2 to 650 cumulative effective batches), with a constant learning rate (`--skip-scheduler`; the declared warmup of 10000 is therefore inert). Milestone indices are not epochs and not proportional to gradient updates. For OpenAI L/14 and B/32 N/U/SSR, run parameter files were located only for steps 6–11 and 7–11 respectively; these are consistent with the other backbones, and steps 1–5 are unresolved.

Table S5: Training settings read from each canonical run's saved parameter record (CUB-200 and NAB-555 training, 100 epochs, seed 0, AdamW-style open_clip trainer with multi-positive contrastive loss, bf16, no extra augmentation, cosine schedule). LR is the base learning rate; LR_{sl} is the learning rate of the structured-routing layers (for N/U it is the unused default). Effective batch is 20 $\times$ gradient-accumulation steps (100 for CUB, 111 for NAB). Settings are parent-specific and not normalized: the OpenAI L/14 and B/32 CUB runs use weight decay 0.2 and warmup 10000, whereas the DFN/MetaCLIP runs use 0.1 and 1000; DFN5B CUB SSR and SSR+ND use $\text{LR}_{\text{sl}} = 5 \times 10^{-6}$ while all other structured runs use 10^{-5}. SSR+ND uses name-branch dropout $p = 0.5$. Deploy ep. is the run's own P1-selected checkpoint.

Backbone	Method	LR	LR_{sl}	WD	Eff. batch	Warmup	Deploy ep.	Routing
CUB-200 training								
DFN5B H/14-378	N	5e-06	1e-05	0.1	2000	1000	40	–
	U	5e-06	1e-05	0.1	2000	1000	25	–
	SSR	5e-06	5e-06	0.1	2000	1000	20	shared-concept

continued on next page

Backbone	Method	LR	LR_{sl}	WD	Eff. batch	Warmup	Deploy ep.	Routing
	SSR+ND	5e-06	5e-06	0.1	2000	1000	60	shared-concept
DFN2B L/14	N	5e-06	1e-05	0.1	2000	1000	20	–
	U	5e-06	1e-05	0.1	2000	1000	20	–
	SSR	5e-06	1e-05	0.1	2000	1000	100	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2000	1000	40	shared-concept
MetaCLIP H/14	N	5e-06	1e-05	0.1	2000	1000	75	–
	U	5e-06	1e-05	0.1	2000	1000	75	–
	SSR	5e-06	1e-05	0.1	2000	1000	100	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2000	1000	75	shared-concept
MetaCLIP L/14	N	5e-06	1e-05	0.1	2000	1000	60	–
	U	5e-06	1e-05	0.1	2000	1000	60	–
	SSR	5e-06	1e-05	0.1	2000	1000	100	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2000	1000	100	shared-concept
OpenAI L/14	N	5e-06	1e-05	0.2	2000	10000	75	–
	U	5e-06	1e-05	0.2	2000	10000	100	–
	SSR	5e-06	1e-05	0.2	2000	10000	90	shared-concept
	SSR+ND	5e-06	1e-05	0.2	2000	10000	90	shared-concept
OpenAI B/32	N	5e-06	1e-05	0.2	2000	10000	100	–
	U	5e-06	1e-05	0.2	2000	10000	60	–
	SSR	5e-06	1e-05	0.2	2000	10000	60	shared-concept
	SSR+ND	5e-06	1e-05	0.2	2000	10000	100	shared-concept
NAB-555 training								
DFN5B H/14-378	N	5e-06	1e-05	0.1	2220	1000	12	–
	U	5e-06	1e-05	0.1	2220	1000	15	–
	SSR	5e-06	1e-05	0.1	2220	1000	40	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2220	1000	75	shared-concept
DFN2B L/14	N	5e-06	1e-05	0.1	2220	1000	50	–
	U	5e-06	1e-05	0.1	2220	1000	35	–
	SSR	5e-06	1e-05	0.1	2220	1000	45	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2220	1000	45	shared-concept
MetaCLIP H/14	N	5e-06	1e-05	0.1	2220	1000	75	–
	U	5e-06	1e-05	0.1	2220	1000	20	–
	SSR	5e-06	1e-05	0.1	2220	1000	35	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2220	1000	35	shared-concept
MetaCLIP L/14	N	5e-06	1e-05	0.1	2220	1000	75	–
	U	5e-06	1e-05	0.1	2220	1000	30	–
	SSR	5e-06	1e-05	0.1	2220	1000	30	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2220	1000	40	shared-concept
OpenAI L/14	N	5e-06	1e-05	0.1	2220	1000	100	–
	U	5e-06	1e-05	0.1	2220	1000	30	–
	SSR	5e-06	1e-05	0.1	2220	1000	100	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2220	1000	45	shared-concept
OpenAI B/32	N	5e-06	1e-05	0.1	2220	1000	100	–
	U	5e-06	1e-05	0.1	2220	1000	60	–
	SSR	5e-06	1e-05	0.1	2220	1000	100	shared-concept
	SSR+ND	5e-06	1e-05	0.1	2220	1000	75	shared-concept

C.3 BACKBONE-SPECIFIC HYPERPARAMETERS

The following deviations from a single recipe were verified in the parameter files and are preserved in the tables: (i) OpenAI L/14 and B/32 CUB runs use weight decay 0.2 and warmup 10000 (DFN/MetaCLIP CUB: 0.1 and 1000; OpenAI NAB: 0.1 and 1000); (ii) DFN5B CUB SSR and SSR+ND use a structured-layer learning rate of 5×10^{-6}, whereas all other CUB/NAB structured

Method	CUB/NAB	ImageNet
N	Ordinary causal text pathway; class-name-only text	same
U	Ordinary causal text pathway; unstructured descriptive text (class name and attributes in one stream)	same
SSR	Routed causal attention over NAME and ATTRIBUTE segments; shared CONCEPT segment (context only, not fused); ATTRIBUTE summary pooled from the <att_end> states; learned-gate Structured Text Fusion of NAME, ATTRIBUTE and EOT summaries (training only); the six added start/end marker embeddings (NAME, ATTRIBUTE, CONCEPT) are randomly initialized and trained	Same routing and fusion; no CONCEPT segment; the four added start/end marker embeddings (NAME, ATTRIBUTE) are randomly initialized and trained
SSR+ND	SSR with NAME-summary dropout $p = 0.5$ at training-time fusion	same
Name-access ablation	FULL: the ATTRIBUTE route reads the full NAME segment; PARTIAL: it reads only the NAME_END token; NO: no direct NAME access (identical to canonical SSR)	–

Table S6: Implementation summary of the training-time routings (reproducibility mapping; all conditions are evaluated with the ordinary text encoder and EOT embedding).

runs use 10^{-5} (ImageNet: 5×10^{-6} throughout); (iii) the canonical DFN5B CUB SSR+ND run differs from the SSR run only by the added name-drop (weight decay 0.1); an earlier historical NAME-drop calibration used weight decay 0.2 and is reported separately in §C.8, but is not used in the canonical main results; (iv) the ImageNet OpenAI N/U/SSR parameter files were located only for steps 6–11 (L/14) and 7–11 (B/32).

C.4 Structured fusion and routing implementation

The gate and AIP parameters are learned during training only; evaluation never invokes fusion (§A.2). For the text-frozen U control used in §G, the entire text encoder including layer norms is locked and only the image encoder and logit scale are trained.

C.5 Checkpoint schedules

ImageNet: 11 chained milestones at cumulative data fractions 0.25, 0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100% (Table S4). CUB/NAB: 100-epoch runs evaluated at dense epochs; the name-access ablation and hard-negative runs were evaluated at epochs 0, 4, 8, 12, 20, 30, 40, 50, 60, 75, 100. The canonical DFN5B/CUB SSR+ND run was evaluated at 8 checkpoints (epochs 0, 4, 8, 12, 20, 60, 75, 100), so its peak-epoch resolution is coarser than for the other methods.

C.6 Multi-seed robustness checks

We report available multi-seed runs as robustness checks rather than treating the six pretrained backbones as stochastic replicates. Some multi-seed runs were conducted under earlier SSR routing variants; where the structured configuration differs from the canonical SSR routing used in the main paper, we report those runs separately and do not pool them with the current trajectory results.

An audit of all archived multi-seed groups found five seeds (10/20/30/40/50) for N, U, and an earlier SSR routing variant, on DFN5B H/14-378 only, for CUB-200, NABirds, and FineView. **No archived group matches the current canonical SSR recipe** (current: shared-concept routing, LR 5×10^{-6}), and there are no multi-seed runs for SSR+ND, ImageNet, or any other backbone; hence there are no "current-recipe replications". All groups are labeled *historical matched robustness checks*. The metric is in-domain validation R@1 at the best checkpoint; P1/P2/P3/CAPP were not evaluated for these runs, so they say nothing about the variance of the current P3/CAPP trajectories. Ordinary in-domain retrieval differences between methods are small relative to, or comparable to, seed-to-seed variation. NABirds runs were walltime-limited and are not converged, and are not an exact stochastic-uncertainty estimate for any result in the paper. Obtaining current-recipe multi-seed numbers would require new runs with the canonical configuration.

Dataset	Method	Seeds	i2t R@1	t2i R@1	Status
CUB-200	N	5	91.98 ± 0.05	94.20 ± 0.06	converged
	U	5	92.02 ± 0.10	94.03 ± 0.05	converged
	earlier SSR routing variant	5	92.01 ± 0.10	94.30 ± 0.11	converged
NABirds	N	5	89.00 ± 0.06	92.78 ± 0.04	not conv.
	U	5	88.91 ± 0.07	92.86 ± 0.24	not conv.
	earlier SSR routing variant	5	88.92 ± 0.09	93.14 ± 0.10	not conv.
FineView	N	5	83.63 ± 0.17	89.42 ± 0.16	converged
	U	5	83.70 ± 0.06	89.00 ± 0.12	converged
	earlier SSR routing variant	5	83.28 ± 0.09	88.91 ± 0.20	converged

Table S7: Archived multi-seed runs (mean $\pm$ sample std; DFN5B H/14-378 only; seeds 10/20/30/40/50; in-domain validation R@1 at the best checkpoint; sample standard deviation). **None of these runs uses the current canonical SSR recipe** (current: shared-concept routing, LR 5×10^{-6}; archived structured runs: earlier routing variants, LR 10^{-5}; the NABirds runs used a different earlier variant from the CUB-200 and FineView runs), so they are historical matched robustness checks, never pooled with the trajectory results, and no SSR+ND, ImageNet, or other-backbone multi-seed runs exist. No P1/P2/P3/CAPP was evaluated for them. NABirds runs were walltime-limited and are not converged (best checkpoints fall at epochs $\approx$19–23, with unequal numbers of logged epochs across seeds). On CUB, historical SSR used 30 accumulation steps versus 100 for N/U, so epoch-based comparisons are confounded.

Dataset	Method	Seed 10	Seed 20	Seed 30	Seed 40	Seed 50
CUB-200	N	91.95/94.22	92.02/94.23	91.92/94.25	92.03/94.10	92.00/94.18
	U	91.93/93.95	92.05/94.08	91.93/94.03	92.00/94.07	92.17/94.03
	hist. SSR	92.08/94.30	91.98/94.18	91.93/94.20	92.15/94.38	91.92/94.43
NABirds	N	89.07/92.76	88.94/92.74	89.02/92.79	89.02/92.84	88.93/92.79
	U	88.92/92.66	89.02/92.74	88.88/92.66	88.86/93.18	88.87/93.04
	hist. SSR	88.78/93.14	88.94/93.31	88.96/93.07	88.90/93.11	89.03/93.05
FineView	N	83.81/89.58	83.65/89.39	83.39/89.17	83.56/89.42	83.77/89.52
	U	83.69/89.15	83.60/88.92	83.75/88.86	83.71/88.98	83.75/89.10
	hist. SSR	83.35/89.14	83.29/88.75	83.11/88.71	83.31/89.11	83.33/88.83

Table S8: Per-seed best-checkpoint i2t / t2i R@1 (%) for the archived multi-seed runs summarized in Table S7.

C.7 COMPUTE AND SOFTWARE NOTES

Each training run used one GPU (one process), 8 CPU cores, and 100 GB memory. Slurm accounting records identify NVIDIA A100 or H100 GPUs for most runs; a minority have untyped GPU allocations from a partition currently comprising B200 nodes, so GPU type varied across runs. Runs used Python 3.10 and a local, untracked open_clip fork reporting version 2.24.0. The current environment contains PyTorch 2.8.0+cu128 with CUDA 12.8; package versions were not logged by individual runs. No code revision identifier is available because the experimental fork was not under version control.

C.8 HISTORICAL NAME-DROP PROBABILITY CALIBRATION

This subsection reports an earlier *historical, single-backbone* calibration of the NAME-summary dropout probability p (DFN5B H/14-378 fine-tuned on CUB-200 or on NABirds-555; $p \in \{0, 0.25, 0.5, 0.75, 1.0\}$; one run per value), evaluated with the P1/P2/P3 definitions of §3. It addresses why $p{=}0.5$ was used and whether the effect of NAME-drop depends on that one value. It is not a main result: it is not pooled with the six-backbone counts, timing statistics, Table 1 medians or canonical trajectories, and, being one run per value, it does not estimate stochastic variance. The runs are matched historical calibration runs that vary p, with the other training settings recorded in their parameter files identical except as follows. In the CUB sweep the $p{=}0$ point is the baseline SSR run (weight decay 0.1, warmup 1000), whereas the four $p > 0$ runs used weight decay 0.2 (warmup 10000, which is inert under fine-tuning from a pretrained checkpoint; 4 rather than 8 data-loader workers; an additional epoch-90 checkpoint), so the $p{=}0$ endpoint is not weight-decay matched. In the NABirds sweep all runs used weight decay 0.1; the $p{=}0$ run was evaluated only through epoch 40 whereas the other four were evaluated to epoch 100.

Target-domain P3 increases with p. Table S9 gives the maximum observed P3 by p. The maximum observed target-domain P3 over the saved checkpoints increases monotonically with p in both sweeps: CUB-200 from 40.42 to 50.35 (+9.93 pp) and NAB-555 from 15.29 to 17.93 (+2.64 pp). The CUB trend is also monotone among the four weight-decay-matched $p > 0$ runs alone (44.60 to 50.35). P1 and P2 remain near their ceiling in this calibration (CUB-trained peak P1 90.37 to 90.01, peak P2 90.16 to 89.97).

Held-out effects are smaller and mixed. Held-out peak P3 changes far less than the in-domain values: CUB-trained to NAB-348 rises from 17.35 to 18.07, CUB-trained to FineView is flat to slightly lower (9.90 to 9.52), NABirds-trained to CUB-59 rises from 53.80 to 56.10, and NABirds-trained to FineView is flat (10.94 to 10.58). We therefore do not read these data as a monotonic transfer benefit of larger p.

Held-out retention. For the CUB-trained sweep, Table S10 gives the relative decrease of held-out P3 from its peak to the P1-selected checkpoint. Moderate dropout ($p{=}0.25$–0.5) has the smallest decrease on both held-out sets (NAB-348 18.8–19.5% versus 23.1% at $p{=}0$ and 20.9% at $p{=}1$; FineView 7.9–9.5% versus 12.4% and 15.4%), a U-shaped pattern in post-peak decrease (equivalently, an inverted-U retention pattern) in this single run per value. The NABirds-trained retention sweep is not reported because its $p{=}0$ run has shorter evaluation coverage (epoch 40 versus 100), which makes the comparison unreliable.

Reading of the calibration. The historical single-backbone calibration shows that increasing NAME-summary dropout strengthens target-domain P3, while held-out retention is best at moderate dropout rather than increasing monotonically. We therefore treat $p{=}0.5$ as a reasonable operating point rather than a globally optimized value.

Relation to the canonical $p{=}0.5$ run. The historical CUB-trained $p{=}0.5$ calibration run used the earlier weight-decay-0.2 recipe, whereas the canonical DFN5B/CUB SSR+ND run uses weight decay 0.1 (it is a one-flag clone of the baseline SSR run). At $p{=}0.5$ the historical and canonical runs show closely matching trajectories on the compared checkpoints (epochs 0, 4, 8, 12, 20, 75 and 100): CUB P1 differs by at most 0.2 pp and CUB P3 by at most 0.27 pp (for example 45.56 versus 45.63 at epoch 20 and 46.82 versus 46.55 at epoch 100), CUB CAPP by at most 0.10 pp (pairwise)

	CUB-trained			NABirds-trained		
p	CUB-200 (target)	NAB-348	FineView	NAB-555 (target)	CUB-59	FineView
0	40.42	17.35	9.90	15.29	53.80	10.94
0.25	44.60	17.34	9.48	16.11	54.17	11.13
0.5	46.82	17.54	9.33	16.85	54.98	11.23
0.75	48.95	17.87	9.48	17.41	55.79	11.15
1	50.35	18.07	9.52	17.93	56.10	10.58

Table S9: Historical NAME-drop probability calibration (DFN5B H/14-378; one run per p): maximum observed P3 (R@1, %) over saved epochs > 0 (the NABirds p=0 run was evaluated only through epoch 40, the others to epoch 100) on the training-domain target and on held-out sets. The p=0.5 CUB-trained run is the earlier weight-decay-0.2 run, not the canonical SSR+ND run. Held-out columns use each run's own P3 peak over epochs (a peak-anywhere convention, not the P1-selected checkpoint).

p	$\rightarrow$ NAB-348 (%)	$\rightarrow$ FineView (%)
0	23.1	12.4
0.25	19.5	7.9
0.5	18.8	9.5
0.75	20.8	10.8
1	20.9	15.4

Table S10: Historical NAME-drop calibration, CUB-trained DFN5B H/14-378: relative held-out P3 decrease, $(\mathrm{P3}_{\mathrm{peak}} - \mathrm{P3}_{\mathrm{P1\text{-}selected}})/\mathrm{P3}_{\mathrm{peak}}$, where the P1-selected checkpoint is each run's own highest-CUB-P1 epoch. Values are those of the archived retention analysis (they agree with recomputation from the peak table's source to rounding). One run per p; the p=0 point is the baseline SSR run.

and 0.14 pp (multi-way), and NAB-348 pairwise CAPP by at most 0.1 pp; P2 and FineView were not compared. The two runs also differ in their inert warmup setting, data-loader workers and one extra checkpoint epoch. This comparison is a robustness bridge indicating that the p=0.5 behavior does not appear to be peculiar to the older weight-decay-0.2 recipe. The two runs are not stochastic replicates, the comparison is not a controlled weight-decay ablation, and it does not show that the full p sweep would be unchanged under weight decay 0.1.

D Capability Acquisition Across Backbones

This section gives the numerical support for Figure 2a–b and §4.1. PT means the pretrained model evaluated directly on the evaluation dataset of each table block, never carried over from the training ontology. ImageNet P3 uses the exact evaluation template "a object with {attributes}". Numbers in parentheses are ImageNet milestone indices (1–11), not epochs.

D.1 ImageNet target trajectories

Table S11 gives, for every backbone and method, the P3 pretrained value, peak, deployment, and post-peak change on the ImageNet-1k validation target. PT is the pretrained checkpoint evaluated on ImageNet val. Differences below roughly 0.5 pp should be read cautiously (§B.7).

Table S11: All six backbones, fine-tuned on ImageNet-1k and evaluated on ImageNet val (target); PT is the pretrained checkpoint evaluated directly on ImageNet val. Deploy is each run's own P1-selected checkpoint and peak is the maximum P3 over the evaluated trajectory; best peak and deployment values are bolded independently within each backbone. Structured supervision produces substantially larger P3 acquisition than N/U, while U often shows an early gain that is reduced by deployment. Numbers in parentheses after P3 peak and P3 deploy give the ImageNet training milestone index (1–11), not an optimizer step count or epoch: milestones are 11 checkpoints at geometrically-spaced cumulative ImageNet-1k data fractions (0.25, 0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100%), with PT as milestone 0. Because each later milestone spans a much larger increment of data than each earlier one, the milestone index is not proportional to the number of gradient updates or to wall-clock training progress. Deploy is each run's own ImageNet-val P1-selected milestone, shared across all evaluation datasets for that run; peak is the milestone of the maximum P3 on that evaluation dataset, which can therefore differ from the deployment milestone and from peak timing on other datasets. When multiple milestones attain the same maximum P3, the earliest is reported.

Method	P3 PT	P3 peak (milestone)	Gain_{peak}	P3 deploy (milestone)	Gain_{deploy}	$\text{Forget}_{post-peak}$
ImageNet val (target)						
N	33.12	33.78 (11)	0.66	33.78 (11)	0.66	0.00
U	33.12	35.38 (2)	2.26	24.80 (11)	-8.32	10.58
SSR	33.12	45.34 (10)	12.22	45.12 (11)	12.00	0.22
SSR+ND	33.12	**48.12 (9)**	**15.00**	**48.00 (11)**	**14.88**	0.12
DFN5B H/14-378						
N	28.60	30.06 (11)	1.46	30.06 (11)	1.46	0.00
U	28.60	29.44 (2)	0.84	28.44 (11)	-0.16	1.00
SSR	28.60	33.90 (10)	5.30	33.56 (11)	4.96	0.34
SSR+ND	28.60	**36.70 (11)**	**8.10**	**36.70 (11)**	**8.10**	0.00
DFN2B L/14						
N	35.94	44.08 (11)	8.14	44.08 (11)	8.14	0.00
U	35.94	38.00 (2)	2.06	32.44 (11)	-3.50	5.56
SSR	35.94	43.04 (11)	7.10	43.04 (11)	7.10	0.00
SSR+ND	35.94	**50.60 (11)**	**14.66**	**50.60 (11)**	**14.66**	0.00
MetaCLIP H/14						
N	31.12	35.86 (8)	4.74	35.48 (11)	4.36	0.38
U	31.12	32.68 (1)	1.56	29.36 (11)	-1.76	3.32
SSR	31.12	37.70 (4)	6.58	36.04 (11)	4.92	1.66
SSR+ND	31.12	**43.86 (11)**	**12.74**	**43.86 (11)**	**12.74**	0.00
MetaCLIP L/14						
N	25.48	30.72 (11)	5.24	30.72 (11)	5.24	0.00
U	25.48	27.76 (11)	2.28	27.76 (11)	2.28	0.00
SSR	25.48	35.14 (11)	9.66	35.14 (11)	9.66	0.00
SSR+ND	25.48	**45.36 (11)**	**19.88**	**45.36 (11)**	**19.88**	0.00
OpenAI L/14						
N	20.44	21.98 (5)	1.54	21.68 (11)	1.24	0.30
U	20.44	20.24 (2)	-0.20	19.92 (11)	-0.52	0.32
SSR	20.44	24.62 (11)	4.18	24.62 (11)	4.18	0.00
SSR+ND	20.44	**29.42 (11)**	**8.98**	**29.42 (11)**	**8.98**	0.00
OpenAI B/32						

D.2 IMAGENET SAME-ONTOLOGY SHIFTS

Table S12 reports the same quantities on ImageNet-V2, ImageNet-R, and ImageNet-Sketch, which share the ImageNet class ontology with the training data. PT is re-evaluated on each shift dataset.

Table S12: All six backbones, fine-tuned on ImageNet-1k and evaluated on the same-ontology shifts ImageNet-V2/-R/-Sketch; PT is the pretrained checkpoint evaluated directly on each listed evaluation dataset. Deploy is each run's own (ImageNet-val) P1-selected checkpoint and peak is the maximum P3 over the evaluated trajectory on that dataset; best values are bolded independently within each backbone. Numbers in parentheses after P3 peak and P3 deploy give the ImageNet training milestone index (1–11), not an optimizer step count or epoch: milestones are 11 checkpoints at geometrically-spaced cumulative ImageNet-1k data fractions (0.25, 0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100%), with PT as milestone 0. Because each later milestone spans a much larger increment of data than each earlier one, the milestone index is not proportional to the number of gradient updates or to wall-clock training progress. Deploy is each run's own ImageNet-val P1-selected milestone, shared across all evaluation datasets for that run; peak is the milestone of the maximum P3 on that evaluation dataset, which can therefore differ from the deployment milestone and from peak timing on other datasets. When multiple milestones attain the same maximum P3, the earliest is reported.

Method	P3 PT	P3 peak (milestone)	Gain_{peak}	P3 deploy (milestone)	Gain_{deploy}	$\text{Forget}_{post-peak}$
ImageNet-V2 (same-ontology shift)						
N	29.29	30.76 (11)	1.47	30.76 (11)	1.47	0.00
U	29.29	30.90 (3)	1.61	22.24 (11)	-7.04	8.66
SSR	29.29	40.68 (10)	11.39	40.48 (11)	11.19	0.20
SSR+ND	29.29	**42.83 (10)**	**13.54**	**42.68 (11)**	**13.39**	0.15
DFN5B H/14-378						
N	24.51	25.85 (11)	1.34	25.85 (11)	1.34	0.00
U	24.51	25.43 (3)	0.92	24.82 (11)	0.31	0.61
SSR	24.51	29.81 (10)	5.30	29.78 (11)	5.27	0.03
SSR+ND	24.51	**32.06 (11)**	**7.56**	**32.06 (11)**	**7.56**	0.00
DFN2B L/14						
N	31.66	38.58 (11)	6.91	38.58 (11)	6.91	0.00
U	31.66	33.53 (2)	1.86	28.50 (11)	-3.17	5.03
SSR	31.66	37.30 (11)	5.64	37.30 (11)	5.64	0.00
SSR+ND	31.66	**44.96 (11)**	**13.30**	**44.96 (11)**	**13.30**	0.00
MetaCLIP H/14						
N	27.39	31.64 (11)	4.25	31.64 (11)	4.25	0.00
U	27.39	30.05 (2)	2.66	25.63 (11)	-1.76	4.42
SSR	27.39	32.92 (4)	5.52	31.74 (11)	4.35	1.17
SSR+ND	27.39	**37.94 (11)**	**10.54**	**37.94 (11)**	**10.54**	0.00
MetaCLIP L/14						
N	22.73	27.46 (11)	4.74	27.46 (11)	4.74	0.00
U	22.73	24.62 (11)	1.89	24.62 (11)	1.89	0.00
SSR	22.73	31.47 (11)	8.75	31.47 (11)	8.75	0.00
SSR+ND	22.73	**39.68 (11)**	**16.95**	**39.68 (11)**	**16.95**	0.00
OpenAI L/14						
N	17.16	18.38 (5)	1.21	18.10 (11)	0.93	0.28
U	17.16	17.54 (2)	0.37	17.00 (11)	-0.16	0.53
SSR	17.16	20.93 (9)	3.77	20.63 (11)	3.47	0.30
SSR+ND	17.16	**24.42 (10)**	**7.25**	**24.20 (11)**	**7.03**	0.22
OpenAI B/32						
ImageNet-R (same-ontology shift)						
N	35.40	43.52 (6)	8.12	42.77 (11)	7.37	0.75
U	35.40	38.52 (3)	3.12	32.33 (11)	-3.07	6.18
SSR	35.40	51.33 (5)	15.93	50.07 (11)	14.67	1.27
SSR+ND	35.40	**53.60 (6)**	**18.20**	**53.13 (11)**	**17.73**	0.47
DFN5B H/14-378						
N	30.62	31.22 (11)	0.60	31.22 (11)	0.60	0.00
U	30.62	32.90 (4)	2.28	31.08 (11)	0.47	1.82
SSR	30.62	35.45 (6)	4.83	34.83 (11)	4.22	0.62
SSR+ND	30.62	**38.50 (11)**	**7.88**	**38.50 (11)**	**7.88**	0.00
DFN2B L/14						
N	42.87	49.25 (10)	6.38	49.20 (11)	6.33	0.05
U	42.87	47.12 (1)	4.25	44.33 (11)	1.47	2.78
SSR	42.87	52.45 (4)	9.58	46.80 (11)	3.93	5.65
SSR+ND	42.87	**58.20 (11)**	**15.33**	**58.20 (11)**	**15.33**	0.00
MetaCLIP H/14						
N	36.35	41.70 (11)	5.35	41.70 (11)	5.35	0.00
U	36.35	41.92 (2)	5.57	38.63 (11)	2.28	3.28
SSR	36.35	46.28 (4)	9.93	40.25 (11)	3.90	6.03
SSR+ND	36.35	**48.65 (5)**	**12.30**	**47.63 (11)**	**11.28**	1.02
MetaCLIP L/14						

continued on next page

Method	P3 PT	P3 peak (milestone)	Gain_{peak}	P3 deploy (milestone)	Gain_{deploy}	$\text{Forget}_{post-peak}$
N	31.45	37.92 (11)	6.47	37.92 (11)	6.47	0.00
U	31.45	37.28 (11)	5.83	37.28 (11)	5.83	0.00
SSR	31.45	42.33 (11)	10.88	42.33 (11)	10.88	0.00
SSR+ND	31.45	**53.08 (11)**	**21.63**	**53.08 (11)**	**21.63**	0.00
OpenAI L/14						
N	22.35	26.08 (5)	3.73	24.13 (11)	1.78	1.95
U	22.35	25.10 (2)	2.75	24.82 (11)	2.47	0.28
SSR	22.35	27.25 (6)	4.90	24.83 (11)	2.48	2.42
SSR+ND	22.35	**29.45 (6)**	**7.10**	**29.08 (11)**	**6.73**	0.37
OpenAI B/32						
ImageNet-Sketch (same-ontology shift)						
N	28.35	30.67 (11)	2.32	30.67 (11)	2.32	0.00
U	28.35	30.55 (3)	2.20	21.37 (11)	-6.98	9.19
SSR	28.35	37.43 (10)	9.08	37.39 (11)	9.04	0.04
SSR+ND	28.35	**39.43 (10)**	**11.08**	**39.40 (11)**	**11.05**	0.03
DFN5B H/14-378						
N	23.85	23.79 (1)	-0.06	23.79 (11)	-0.06	0.00
U	23.85	24.56 (3)	0.71	23.40 (11)	-0.45	1.16
SSR	23.85	27.70 (11)	3.85	27.70 (11)	3.85	0.00
SSR+ND	23.85	**29.25 (11)**	**5.40**	**29.25 (11)**	**5.40**	0.00
DFN2B L/14						
N	29.31	36.24 (11)	6.93	36.24 (11)	6.93	0.00
U	29.31	32.42 (2)	3.11	28.44 (11)	-0.88	3.98
SSR	29.31	34.52 (5)	5.21	33.15 (11)	3.83	1.38
SSR+ND	29.31	**38.80 (11)**	**9.49**	**38.80 (11)**	**9.49**	0.00
MetaCLIP H/14						
N	25.44	28.92 (11)	3.47	28.92 (11)	3.47	0.00
U	25.44	29.06 (2)	3.61	24.72 (11)	-0.72	4.34
SSR	25.44	29.92 (4)	4.48	28.32 (11)	2.87	1.60
SSR+ND	25.44	**31.74 (11)**	**6.30**	**31.74 (11)**	**6.30**	0.00
MetaCLIP L/14						
N	19.41	24.96 (11)	5.55	24.96 (11)	5.55	0.00
U	19.41	21.86 (11)	2.46	21.86 (11)	2.46	0.00
SSR	19.41	26.72 (11)	7.31	26.72 (11)	7.31	0.00
SSR+ND	19.41	**31.88 (11)**	**12.47**	**31.88 (11)**	**12.47**	0.00
OpenAI L/14						
N	12.70	15.08 (5)	2.38	14.30 (11)	1.60	0.78
U	12.70	14.07 (2)	1.37	13.10 (11)	0.40	0.97
SSR	12.70	15.07 (6)	2.37	14.72 (11)	2.02	0.35
SSR+ND	12.70	**17.25 (11)**	**4.56**	**17.25 (11)**	**4.56**	0.00
OpenAI B/32						

D.3 IMAGENET-TRAINED CROSS-ONTOLOGY TRAJECTORIES

Table S13 evaluates the ImageNet-trained models on CUB-200, NAB-348 (strict dataset-class-disjoint held-out), and FineView Butterfly. PT is re-evaluated on each held-out dataset; it is not carried over from ImageNet.

Table S13: All six backbones, fine-tuned on ImageNet-1k and evaluated cross-ontology on CUB-200, NAB-348 (strict dataset-class-disjoint held-out), and FineView Butterfly (FVB); PT is the pretrained checkpoint evaluated directly on each held-out dataset. Deploy is each run's own P1-selected checkpoint (shared with Table S11) and peak is the maximum P3 over the trajectory; best peak and deployment values are bolded independently within each backbone (trajectory regimes in Table S16). Cross-ontology P3 acquisition is substantially smaller and more backbone-dependent than acquisition on ImageNet and its same-ontology shifts, and its peak often occurs earlier than the shared deployment milestone. Numbers in parentheses after P3 peak and P3 deploy give the ImageNet training milestone index (1–11), not an optimizer step count or epoch: milestones are 11 checkpoints at geometrically-spaced cumulative ImageNet-1k data fractions (0.25, 0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100%), with PT as milestone 0. Because each later milestone spans a much larger increment of data than each earlier one, the milestone index is not proportional to the number of gradient updates or to wall-clock training progress. Deploy is each run's own ImageNet-val P1-selected milestone, shared across all evaluation datasets for that run; peak is the milestone of the maximum P3 on that evaluation dataset, which can therefore differ from the deployment milestone and from peak timing on other datasets. When multiple milestones attain the same maximum P3, the earliest is reported.

Method	P3 PT	P3 peak (milestone)	Gain_{peak}	P3 deploy (milestone)	Gain_{deploy}	$\text{Forget}_{post-peak}$
CUB-200 (cross-ontology held-out)						
N	34.36	33.64 (1)	-0.72	20.73 (11)	-13.63	12.91
U	34.36	**35.48 (1)**	**1.12**	15.26 (11)	-19.11	20.23
SSR	34.36	34.76 (2)	0.40	23.35 (11)	-11.01	11.41
SSR+ND	34.36	34.88 (3)	0.52	**24.99 (11)**	**-9.37**	9.89
DFN5B H/14-378						
N	27.96	27.96 (1)	0.00	25.92 (11)	-2.04	2.04
U	27.96	27.65 (1)	-0.31	20.14 (11)	-7.82	7.51
SSR	27.96	28.15 (5)	0.19	26.01 (11)	-1.95	2.14
SSR+ND	27.96	**28.27 (1)**	**0.31**	**26.92 (11)**	**-1.04**	1.35
DFN2B L/14						
N	28.22	30.57 (6)	2.35	27.74 (11)	-0.48	2.83
U	28.22	23.16 (1)	-5.06	17.09 (11)	-11.13	6.08
SSR	28.22	30.70 (3)	2.49	26.79 (11)	-1.43	3.92
SSR+ND	28.22	**31.27 (3)**	**3.05**	**29.50 (11)**	**1.28**	1.78
MetaCLIP H/14						
N	23.90	24.84 (7)	0.93	24.70 (11)	0.79	0.14
U	23.90	19.78 (1)	-4.12	18.57 (11)	-5.33	1.21
SSR	23.90	25.51 (3)	1.61	22.35 (11)	-1.55	3.16
SSR+ND	23.90	**28.29 (10)**	**4.38**	**27.17 (11)**	**3.26**	1.12
MetaCLIP L/14						
N	16.71	17.41 (1)	0.71	16.19 (11)	-0.52	1.23
U	16.71	16.64 (11)	-0.07	16.64 (11)	-0.07	0.00
SSR	16.71	18.97 (3)	2.26	17.26 (11)	0.55	1.71
SSR+ND	16.71	**20.54 (8)**	**3.83**	**19.35 (11)**	**2.64**	1.19
OpenAI L/14						
N	13.76	13.53 (1)	-0.22	12.46 (11)	-1.29	1.07
U	13.76	12.58 (1)	-1.17	11.81 (11)	-1.95	0.78
SSR	13.76	**15.53 (7)**	**1.78**	**14.70 (11)**	**0.95**	0.83
SSR+ND	13.76	15.22 (8)	1.47	14.27 (11)	0.52	0.95
OpenAI B/32						
NAB-348 (strict cross-ontology held-out)						
N	13.81	13.41 (1)	-0.41	9.25 (11)	-4.56	4.15
U	13.81	14.28 (1)	0.47	4.61 (11)	-9.21	9.67
SSR	13.81	14.68 (6)	0.87	11.25 (11)	-2.56	3.43
SSR+ND	13.81	**15.04 (6)**	**1.23**	**12.22 (11)**	**-1.59**	2.82
DFN5B H/14-378						
N	8.85	8.84 (1)	-0.01	7.07 (11)	-1.78	1.76
U	8.85	8.65 (1)	-0.20	4.09 (11)	-4.76	4.56
SSR	8.85	8.97 (2)	0.11	**8.17 (11)**	**-0.68**	0.80
SSR+ND	8.85	**9.26 (4)**	**0.41**	7.44 (11)	-1.41	1.82
DFN2B L/14						
N	11.59	**12.32 (5)**	**0.72**	**10.75 (11)**	**-0.84**	1.57
U	11.59	8.49 (1)	-3.11	3.45 (11)	-8.15	5.04
SSR	11.59	12.06 (2)	0.47	10.08 (11)	-1.52	1.99
SSR+ND	11.59	12.17 (1)	0.58	9.79 (11)	-1.81	2.38
MetaCLIP H/14						

continued on next page

Method	P3 PT	P3 peak (milestone)	Gain_{peak}	P3 deploy (milestone)	Gain_{deploy}	$\text{Forget}_{post-peak}$
N	7.02	6.76 (1)	-0.25	**6.25 (11)**	**-0.77**	0.52
U	7.02	4.20 (1)	-2.82	3.70 (11)	-3.32	0.50
SSR	7.02	7.04 (2)	0.02	4.17 (11)	-2.85	2.87
SSR+ND	7.02	**7.38 (1)**	**0.36**	5.78 (11)	-1.24	1.60
MetaCLIP L/14						
N	7.41	7.57 (2)	0.16	5.80 (11)	-1.60	1.77
U	7.41	5.45 (1)	-1.96	4.14 (11)	-3.27	1.30
SSR	7.41	8.05 (3)	0.64	6.70 (11)	-0.71	1.35
SSR+ND	7.41	**8.40 (8)**	**0.99**	**7.56 (11)**	**0.15**	0.84
OpenAI L/14						
N	4.23	4.12 (1)	-0.11	3.48 (11)	-0.75	0.64
U	4.23	3.37 (10)	-0.87	3.18 (11)	-1.05	0.18
SSR	4.23	4.89 (7)	0.66	4.62 (11)	0.39	0.27
SSR+ND	4.23	**4.96 (11)**	**0.73**	**4.96 (11)**	**0.73**	0.00
OpenAI B/32						
FineView Butterfly (FVB) (cross-ontology held-out)						
N	8.52	8.42 (1)	-0.10	6.42 (11)	-2.10	2.00
U	8.52	8.75 (3)	0.23	4.06 (11)	-4.46	4.69
SSR	8.52	**8.85 (6)**	**0.33**	**7.13 (11)**	**-1.38**	1.71
SSR+ND	8.52	8.38 (1)	-0.13	6.77 (11)	-1.75	1.62
DFN5B H/14-378						
N	6.67	6.85 (11)	0.17	6.85 (11)	0.17	0.00
U	6.67	6.44 (1)	-0.23	5.87 (11)	-0.81	0.58
SSR	6.67	**7.71 (11)**	**1.04**	**7.71 (11)**	**1.04**	0.00
SSR+ND	6.67	7.44 (11)	0.77	7.44 (11)	0.77	0.00
DFN2B L/14						
N	6.52	6.77 (5)	0.25	6.19 (11)	-0.33	0.58
U	6.52	5.48 (1)	-1.04	5.00 (11)	-1.52	0.48
SSR	6.52	**7.13 (6)**	**0.62**	5.98 (11)	-0.54	1.15
SSR+ND	6.52	7.04 (11)	0.52	**7.04 (11)**	**0.52**	0.00
MetaCLIP H/14						
N	4.73	5.10 (11)	0.37	5.10 (11)	0.37	0.00
U	4.73	4.71 (7)	-0.02	4.44 (11)	-0.29	0.27
SSR	4.73	5.13 (8)	0.40	4.73 (11)	0.00	0.40
SSR+ND	4.73	**5.58 (10)**	**0.85**	**5.31 (11)**	**0.58**	0.27
MetaCLIP L/14						
N	3.75	4.13 (5)	0.38	3.25 (11)	-0.50	0.88
U	3.75	4.12 (7)	0.37	3.42 (11)	-0.33	0.69
SSR	3.75	**5.25 (5)**	**1.50**	3.62 (11)	-0.13	1.63
SSR+ND	3.75	4.23 (7)	0.48	**3.77 (11)**	**0.02**	0.46
OpenAI L/14						
N	2.73	3.63 (4)	0.90	3.13 (11)	0.40	0.50
U	2.73	3.23 (2)	0.50	2.62 (11)	-0.12	0.62
SSR	2.73	4.02 (9)	1.29	3.44 (11)	0.71	0.58
SSR+ND	2.73	**4.04 (9)**	**1.31**	**3.87 (11)**	**1.13**	0.17
OpenAI B/32						

D.4 Dense representative CUB→NAB trajectory

Table S14 gives the epoch-level summary underlying Figure 2a for the representative DFN5B H/14-378 CUB-trained runs. This is a single representative backbone, not a six-backbone result.

Table S14: Representative DFN5B H/14-378 fine-tuned on CUB-200, with dense trajectory evaluation on CUB-200 (target) and NAB-348 (strict cross-ontology held-out); PT is the pretrained checkpoint evaluated directly on each dataset. Peak is the maximum P3 over the evaluated trajectory and deploy is each method's own P1-selected checkpoint; best peak and deployment values are bolded independently within each dataset. All four fine-tuning conditions acquire held-out P3 above the pretrained state at some point, but their peak timing and post-peak retention differ substantially.

Method	PT	Peak	Peak ep.	Deploy	Deploy ep.	Gain_{peak}	Gain_{deploy}	$\text{Forget}_{post-peak}$
CUB-200 (target)								
N	34.35	35.23	4	31.12	40	0.88	-3.23	4.11
U	34.45	38.87	12	35.78	25	4.42	1.33	3.09
SSR	34.35	40.42	20	40.42	20	6.08	6.08	0.00
SSR+ND	34.35	**46.55**	75	**46.31**	60	**12.20**	**11.96**	0.24
NAB-348 (strict cross-ontology held-out)								
N	13.85	15.59	12	13.74	40	1.74	-0.11	1.85
U	13.87	16.38	10	15.14	25	2.51	1.27	1.24
SSR	13.85	17.32	12	**16.90**	20	3.48	**3.06**	0.42
SSR+ND	13.85	**17.66**	20	14.50	60	**3.81**	0.65	3.16

D.5 Six-backbone acquisition summary

Table S15 summarizes peak P3 gain across the six pretrained backbones. The backbones are heterogeneous and are not stochastic replicates; means and medians are descriptive.

Table S15: Six-backbone descriptive summary of peak P3 gain over pretrained, fine-tuned on ImageNet-1k and evaluated on ImageNet val (target), its same-ontology shifts, and cross-ontology held-out data (CUB-200, NAB-348, FineView Butterfly); PT for each backbone is evaluated directly on the corresponding dataset. Means and medians are descriptive across the six deliberately diverse pretrained backbones, not stochastic-replicate estimates; best values within each distribution are bolded, with ties retained. Peak P3 gains are largest under structured supervision on ImageNet-family evaluations, whereas cross-ontology gains are markedly smaller.

Method	Mean Gain_{peak} (pp)	Median Gain_{peak} (pp)	N positive	N >0.5pp
ImageNet val (target)				
N	3.63	3.14	**6/6**	**6/6**
U	1.47	1.81	5/6	5/6
SSR	7.51	6.84	**6/6**	**6/6**
SSR+ND	**13.23**	**13.70**	**6/6**	**6/6**
ImageNet-V2 (same-ontology shift)				
N	3.32	2.86	**6/6**	**6/6**
U	1.55	1.74	**6/6**	5/6
SSR	6.73	5.58	**6/6**	**6/6**
SSR+ND	**11.52**	**11.92**	**6/6**	**6/6**
ImageNet-R (same-ontology shift)				
N	5.11	5.87	**6/6**	**6/6**
U	3.97	3.68	**6/6**	**6/6**
SSR	9.34	9.76	**6/6**	**6/6**
SSR+ND	**13.74**	**13.82**	**6/6**	**6/6**

continued on next page

Method	Mean Gain$_{peak}$ (pp)	Median Gain$_{peak}$ (pp)	N positive	N >0.5pp
ImageNet-Sketch (same-ontology shift)				
N	3.43	2.93	5/6	5/6
U	2.24	2.33	**6/6**	**6/6**
SSR	5.38	4.84	**6/6**	**6/6**
SSR+ND	**8.21**	**7.89**	**6/6**	**6/6**
CUB-200 (cross-ontology)				
N	0.51	0.35	3/6	3/6
U	-1.60	-0.74	1/6	1/6
SSR	1.45	1.69	**6/6**	4/6
SSR+ND	**2.26**	**2.26**	**6/6**	**5/6**
NAB-348 (cross-ontology, strict held-out)				
N	0.02	-0.06	2/6	1/6
U	-1.41	-1.42	1/6	0/6
SSR	0.46	0.55	**6/6**	3/6
SSR+ND	**0.72**	**0.65**	**6/6**	**4/6**
FineView Butterfly (FVB) (cross-ontology)				
N	0.33	0.31	5/6	1/6
U	-0.03	0.11	3/6	0/6
SSR	**0.86**	**0.83**	**6/6**	**4/6**
SSR+ND	0.63	0.64	5/6	**4/6**

D.6 TRAJECTORY-REGIME SUMMARY

Table S16 classifies each cross-ontology cell by whether peak gain was retained, partially reduced, lost, or never clearly acquired; Table S17 lists the pretrained backbones.

Table S16: All six ImageNet-1k-trained backbones, evaluated cross-ontology on CUB-200, NAB-348 (strict held-out), and FineView Butterfly (FVB); PT is the pretrained checkpoint evaluated directly on the relevant held-out dataset. Each cell reports peak P3 gain over pretrained (pp) and a trajectory regime describing whether that gain was retained, partially reduced, or fully lost by deployment, or never clearly acquired. Poor deployed transfer can arise either because a capability was never clearly acquired or because it was acquired earlier and subsequently reduced.

Backbone	Method	CUB-200	NAB-348	FVB
DFN5B H/14-378	N	-0.72, no acquisition / degraded	-0.41, no acquisition / degraded	-0.10, no acquisition / degraded
	U	1.12, acquire→lose	0.47, acquire→lose	0.23, acquire→lose
	SSR	0.40, acquire→lose	0.87, acquire→lose	0.33, acquire→lose
	SSR+ND	0.52, acquire→lose	1.23, acquire→lose	-0.14, no acquisition / degraded
DFN2B L/14	N	0.00, no acquisition / degraded	-0.01, no acquisition / degraded	0.17, acquire→retain
	U	-0.31, no acquisition / degraded	-0.20, no acquisition / degraded	-0.23, no acquisition / degraded
	SSR	0.19, acquire→lose	0.12, acquire→lose	1.04, acquire→retain
	SSR+ND	0.31, acquire→lose	0.41, acquire→lose	0.77, acquire→retain
MetaCLIP H/14	N	2.35, acquire→lose	0.72, acquire→lose	0.25, acquire→lose
	U	-5.06, no acquisition / degraded	-3.11, no acquisition / degraded	-1.04, no acquisition / degraded
	SSR	2.48, acquire→lose	0.47, acquire→lose	0.62, acquire→lose
	SSR+ND	3.06, acquire→partial decline	0.58, acquire→lose	0.52, acquire→retain
MetaCLIP L/14	N	0.93, acquire→partial decline	-0.25, no acquisition / degraded	0.36, acquire→retain
	U	-4.12, no acquisition / degraded	-2.82, no acquisition / degraded	-0.02, no acquisition / degraded
	SSR	1.60, acquire→lose	0.02, no acquisition / degraded	0.40, acquire→lose
	SSR+ND	4.38, acquire→partial decline	0.36, acquire→lose	0.85, acquire→partial decline
OpenAI L/14	N	0.71, acquire→lose	0.16, acquire→lose	0.38, acquire→lose
	U	-0.07, no clear acquisition	-1.96, no acquisition / degraded	0.36, acquire→lose
	SSR	2.26, acquire→partial decline	0.64, acquire→lose	1.50, acquire→lose
	SSR+ND	3.83, acquire→partial decline	0.99, acquire→partial decline	0.48, acquire→lose

continued on next page

Backbone	Method	CUB-200	NAB-348	FVB
OpenAI B/32	N	-0.22, no acquisition / degraded	-0.12, no acquisition / degraded	0.90, acquire→partial decline
	U	-1.17, no acquisition / degraded	-0.87, no acquisition / degraded	0.50, acquire→lose
	SSR	1.78, acquire→partial decline	0.66, acquire→lose	1.29, acquire→partial decline
	SSR+ND	1.47, acquire→partial decline	0.73, acquire→retain	1.31, acquire→partial decline

Table S17: The six pretrained backbones used throughout the ImageNet-1k experiments (Figure 2b), listed before any downstream fine-tuning. The final column is pretrained P1 (class-name retrieval R@1) evaluated on ImageNet val; no downstream trajectory result is represented in this table.

Backbone	Family	Params	Pretraining source	ImageNet-val P1 (%)
DFN5B H/14-378	DFN	986M	DFN5B (5B pool)	83.9
DFN2B L/14	DFN	427M	DFN-2B (2B pool)	81.8
MetaCLIP H/14	MetaCLIP	986M	MetaCLIP (fullcc)	80.3
MetaCLIP L/14	MetaCLIP	427M	MetaCLIP (fullcc)	79.0
OpenAI L/14	OpenAI	427M	OpenAI (WIT-400M)	73.3
OpenAI B/32	OpenAI	151M	OpenAI (WIT-400M)	63.1

D.7 P1 CONTEXT

Table S18 reports deployed class-name retrieval (P1) next to P3 peak, to show that large differences in P3 acquisition occur while deployed P1 is comparatively similar.

Table S18: All six backbones, fine-tuned on ImageNet-1k and evaluated on ImageNet val (target) and its same-ontology shifts; PT is the pretrained checkpoint evaluated directly on each listed dataset. P1 deploy is class-name retrieval at each run's own P1-selected deployment checkpoint, and P3 peak is the maximum P3 over the trajectory; within each backbone, the highest deployed P1 and peak P3 are bolded independently. Large differences in P3 acquisition can occur while deployed P1 remains comparatively similar.

Method	P1 PT	P1 deploy	P3 PT	P3 peak	Gain$_{peak}$(P3)
ImageNet val (target)					
N	83.90	**87.82**	33.12	33.78	0.66
U	83.90	87.66	33.12	35.38	2.26
SSR	83.90	87.00	33.12	45.34	12.22
SSR+ND	83.90	86.96	33.12	**48.12**	15.00
DFN5B H/14-378					
N	81.84	**85.98**	28.60	30.06	1.46
U	81.84	85.66	28.60	29.44	0.84
SSR	81.84	85.00	28.60	33.90	5.30
SSR+ND	81.84	84.86	28.60	**36.70**	8.10
DFN2B L/14					
N	80.26	86.32	35.94	44.08	8.14
U	80.26	**86.74**	35.94	38.00	2.06
SSR	80.26	85.84	35.94	43.04	7.10
SSR+ND	80.26	85.42	35.94	**50.60**	14.66
MetaCLIP H/14					

continued on next page

Method	P1 PT	P1 deploy	P3 PT	P3 peak	Gain$_{peak}$(P3)
N	79.04	85.66	31.12	35.86	4.74
U	79.04	**85.84**	31.12	32.68	1.56
SSR	79.04	84.86	31.12	37.70	6.58
SSR+ND	79.04	84.54	31.12	**43.86**	12.74
MetaCLIP L/14					
N	73.32	**84.98**	25.48	30.72	5.24
U	73.32	84.66	25.48	27.76	2.28
SSR	73.32	83.62	25.48	35.14	9.66
SSR+ND	73.32	83.30	25.48	**45.36**	19.88
OpenAI L/14					
N	63.08	**75.42**	20.44	21.98	1.54
U	63.08	75.16	20.44	20.24	-0.20
SSR	63.08	71.26	20.44	24.62	4.18
SSR+ND	63.08	72.00	20.44	**29.42**	8.98
OpenAI B/32					
ImageNet-V2 (same-ontology shift)					
N	77.98	81.40	29.29	30.76	1.47
U	77.98	**81.44**	29.29	30.90	1.61
SSR	77.98	80.24	29.29	40.68	11.39
SSR+ND	77.98	80.51	29.29	**42.83**	13.54
DFN5B H/14-378					
N	75.10	78.76	24.51	25.85	1.34
U	75.10	**79.00**	24.51	25.43	0.92
SSR	75.10	78.50	24.51	29.81	5.30
SSR+ND	75.10	78.07	24.51	**32.06**	7.55
DFN2B L/14					
N	73.06	79.10	31.66	38.58	6.91
U	73.06	**79.11**	31.66	33.53	1.86
SSR	73.06	78.44	31.66	37.30	5.64
SSR+ND	73.06	78.65	31.66	**44.96**	13.30
MetaCLIP H/14					
N	72.20	78.04	27.39	31.64	4.25
U	72.20	**78.67**	27.39	30.05	2.65
SSR	72.20	77.31	27.39	32.92	5.52
SSR+ND	72.20	77.69	27.39	**37.94**	10.54
MetaCLIP L/14					
N	67.31	**77.55**	22.73	27.46	4.74
U	67.31	77.41	22.73	24.62	1.89
SSR	67.31	76.32	22.73	31.47	8.75
SSR+ND	67.31	76.22	22.73	**39.68**	16.95
OpenAI L/14					
N	54.73	**64.70**	17.16	18.38	1.21
U	54.73	64.43	17.16	17.54	0.37
SSR	54.73	61.14	17.16	20.93	3.77
SSR+ND	54.73	61.16	17.16	**24.42**	7.25
OpenAI B/32					
ImageNet-R (same-ontology shift)					
N	91.50	89.72	35.40	43.52	8.12
U	91.50	89.82	35.40	38.52	3.12
SSR	91.50	89.73	35.40	51.33	15.93
SSR+ND	91.50	**90.20**	35.40	**53.60**	18.20
DFN5B H/14-378					

continued on next page

Method	P1 PT	P1 deploy	P3 PT	P3 peak	Gain$_{peak}$(P3)
N	88.77	88.42	30.62	31.22	0.60
U	88.77	**88.95**	30.62	32.90	2.28
SSR	88.77	88.12	30.62	35.45	4.83
SSR+ND	88.77	88.02	30.62	**38.50**	7.88
DFN2B L/14					
N	90.33	90.58	42.87	49.25	6.38
U	90.33	**91.30**	42.87	47.12	4.25
SSR	90.33	90.13	42.87	52.45	9.58
SSR+ND	90.33	90.87	42.87	**58.20**	15.33
MetaCLIP H/14					
N	88.93	89.03	36.35	41.70	5.35
U	88.93	**89.60**	36.35	41.92	5.57
SSR	88.93	87.88	36.35	46.28	9.93
SSR+ND	88.93	88.88	36.35	**48.65**	12.30
MetaCLIP L/14					
N	83.47	86.00	31.45	37.92	6.47
U	83.47	**86.77**	31.45	37.28	5.83
SSR	83.47	85.45	31.45	42.33	10.88
SSR+ND	83.47	86.17	31.45	**53.08**	21.63
OpenAI L/14					
N	65.03	64.77	22.35	26.08	3.73
U	65.03	**66.07**	22.35	25.10	2.75
SSR	65.03	62.60	22.35	27.25	4.90
SSR+ND	65.03	63.53	22.35	**29.45**	7.10
OpenAI B/32					
ImageNet-Sketch (same-ontology shift)					
N	72.45	71.68	28.35	30.67	2.32
U	72.45	70.54	28.35	30.55	2.20
SSR	72.45	70.84	28.35	37.43	9.08
SSR+ND	72.45	**71.74**	28.35	**39.43**	11.08
DFN5B H/14-378					
N	68.64	68.43	23.85	23.79	-0.06
U	68.64	**68.78**	23.85	24.56	0.71
SSR	68.64	68.64	23.85	27.70	3.85
SSR+ND	68.64	68.68	23.85	**29.25**	5.40
DFN2B L/14					
N	69.53	69.52	29.31	36.24	6.93
U	69.53	**69.61**	29.31	32.42	3.11
SSR	69.53	69.11	29.31	34.52	5.21
SSR+ND	69.53	69.44	29.31	**38.80**	9.49
MetaCLIP H/14					
N	67.51	67.50	25.44	28.92	3.47
U	67.51	**67.56**	25.44	29.06	3.61
SSR	67.51	66.47	25.44	29.92	4.48
SSR+ND	67.51	66.99	25.44	**31.74**	6.30
MetaCLIP L/14					
N	57.34	**61.36**	19.41	24.96	5.55
U	57.34	61.16	19.41	21.86	2.46
SSR	57.34	60.51	19.41	26.72	7.32
SSR+ND	57.34	60.47	19.41	**31.88**	12.47
OpenAI L/14					

continued on next page

Method	P1 PT	P1 deploy	P3 PT	P3 peak	Gain$_{peak}$(P3)
N	41.18	42.57	12.70	15.08	2.38
U	41.18	**42.65**	12.70	14.07	1.37
SSR	41.18	39.81	12.70	15.07	2.37
SSR+ND	41.18	40.63	12.70	**17.25**	4.56
OpenAI B/32					

E SUPERVISION DETERMINES THE ACQUIRED CAPABILITY

This section supports Table 1, Figure 3, and §4.2. The CAPP tables report all six backbones (the DFN/MetaCLIP backbones plus the additional OpenAI L/14 and B/32 backbones, marked with a divider) under the same matched protocol; the six are not pooled as replicates. CUB target structured improvements are consistent, whereas NAB-target and held-out effects are backbone-dependent, and held-out CAPP gains are not universal.

E.1 DEPLOYMENT P1/P2/P3

Table S19 gives P1, P2, and P3 at each run's own P1-selected deployment checkpoint on ImageNet val for all six backbones.

Table S19: All six backbones, fine-tuned on ImageNet-1k and evaluated on ImageNet val (target); PT is the pretrained checkpoint evaluated on ImageNet val. P1/P2/P3 (R@1 %) are all measured at each run's own P1-selected *deployment* checkpoint, not peak P3 (underlies Table 1). U produces the strongest name-containing P2 behavior, whereas structured SSR and especially SSR+ND yield substantially stronger name-free P3 despite broadly similar (not identical) P1.

DFN5B H/14-378

Method	P1	P2	P3
PT	83.90	79.48	33.12
N	87.82	86.54	33.78
U	87.66	87.90	24.80
SSR	87.00	86.10	45.12
SSR+ND	86.96	86.12	48.00

DFN2B L/14

Method	P1	P2	P3
PT	81.84	76.32	28.60
N	85.98	79.66	30.06
U	85.66	85.98	28.44
SSR	85.00	81.86	33.56
SSR+ND	84.86	81.98	36.70

MetaCLIP H/14

Method	P1	P2	P3
PT	80.26	75.56	35.94
N	86.32	83.24	44.08
U	86.74	86.64	32.44
SSR	85.84	83.96	43.04
SSR+ND	85.42	84.38	50.60

MetaCLIP L/14

Method	P1	P2	P3
PT	79.04	72.70	31.12
N	85.66	80.84	35.48
U	85.84	86.04	29.36
SSR	84.86	81.88	36.04
SSR+ND	84.54	82.26	43.86

OpenAI L/14

Method	P1	P2	P3
PT	73.32	69.70	25.48
N	84.98	81.36	30.72
U	84.66	85.18	27.76
SSR	83.62	82.14	35.14
SSR+ND	83.30	82.48	45.36

OpenAI B/32

Method	P1	P2	P3
PT	63.08	58.32	20.44
N	75.42	68.80	21.68
U	75.16	75.82	19.92
SSR	71.26	69.24	24.62
SSR+ND	72.00	70.48	29.42

E.2 P2 NAME-SWAP AND ATTRIBUTE-SWAP DIAGNOSTIC

Table S20 reports ordinary, name-swapped, and attribute-swapped P2 with class-level bootstrap intervals (§B.7). Ordinary P2 remains overwhelmingly class-name anchored under every condition, including PT; SSR does not remove this name reliance.

Table S20: DFN5B H/14-378 fine-tuned on CUB-200, evaluated on CUB-200 (target) and NAB-348 (strict cross-ontology held-out) under five conditions (PT, N, U, SSR, SSR+ND); PT is the pretrained checkpoint evaluated directly on each listed dataset, and all fine-tuned methods are at their own P1-selected deployment checkpoint. Ordinary, name-swapped, and attribute-swapped P2 (R@1 %) are shown with Δ_{name}/Δ_{attr}, the pp change from ordinary, with class-level bootstrap 95% CIs (1000 replicates). Name-swap nearly eliminates P2 retrieval, whereas attribute-swap preserves most performance, showing that ordinary name+attribute retrieval remains strongly class-name anchored.

Method	P2 ord.	P2 name-sw.	P2 attr-sw.	Δ_{name} [95% CI]	Δ_{attr} [95% CI]	P3
CUB-200 (target)						
PT	86.31	0.10	76.54	-86.2 [-88.6, -83.4]	-9.8 [-13.2, -6.6]	34.36
N	90.68	0.07	89.68	-90.6 [-92.1, -88.8]	-1.0 [-1.9, -0.2]	31.10
U	91.35	0.05	90.63	-91.3 [-92.8, -89.7]	-0.7 [-1.5, -0.1]	35.78
SSR	89.44	0.14	85.33	-89.3 [-91.3, -86.9]	-4.1 [-6.6, -2.0]	40.40
SSR+ND	90.04	0.22	86.92	-89.8 [-91.5, -87.8]	-3.1 [-4.7, -1.7]	46.31
NAB-348 (strict held-out)						
PT	82.78	0.20	74.67	-82.6 [-84.6, -80.3]	-8.1 [-10.4, -5.9]	13.86
N	73.14	0.40	64.41	-72.7 [-75.8, -69.8]	-8.7 [-11.6, -6.0]	13.76
U	78.09	0.11	73.28	-78.0 [-80.6, -75.2]	-4.8 [-6.6, -3.1]	15.14
SSR	78.98	0.33	68.39	-78.6 [-81.0, -76.0]	-10.6 [-13.5, -7.8]	16.82
SSR+ND	73.00	0.40	63.31	-72.6 [-75.3, -69.6]	-9.7 [-12.6, -6.7]	14.54

E.3 SSR VERSUS SSR+ND TARGET P3 ACQUISITION

Table S21 lists all 18 backbone×training-domain pairs.

Table S21: All six backbones, fine-tuned separately on ImageNet-1k, CUB-200, and NABirds (18 backbone×training-domain pairs), each evaluated on its own target training domain (ImageNet-trained → ImageNet val, CUB-trained → CUB-200, NAB-trained → NAB-555) – this is a target-domain comparison, not a held-out-transfer table. Gain_{peak} is target P3 peak minus target PT P3 (PT evaluated directly on the target dataset), Gain_{deploy} uses each method's own P1-selected deployment checkpoint, and Forget is their difference (underlies Figure 3b; all values pp). SSR+ND exceeds SSR in peak target P3 acquisition in all 18 tested backbone × training-domain pairs.

		SSR			SSR+ND		
Backbone	Train	Gain_{peak}	Gain_{deploy}	Forget	Gain_{peak}	Gain_{deploy}	Forget
DFN5B H/14-378	ImageNet-trained	12.22	12.00	0.22	**15.00**	14.88	0.12
DFN5B H/14-378	CUB-trained	6.08	6.08	0.00	**12.20**	11.96	0.24
DFN5B H/14-378	NAB-trained	4.09	-0.23	4.32	**5.65**	1.02	4.62
DFN2B L/14	ImageNet-trained	5.30	4.96	0.34	**8.10**	8.10	0.00
DFN2B L/14	CUB-trained	3.78	3.69	0.09	**8.99**	8.99	0.00
DFN2B L/14	NAB-trained	1.01	0.58	0.42	**1.56**	1.54	0.03
MetaCLIP H/14	ImageNet-trained	7.10	7.10	0.00	**14.66**	14.38	0.28
MetaCLIP H/14	CUB-trained	5.70	5.63	0.07	**13.77**	13.77	0.00
MetaCLIP H/14	NAB-trained	3.09	2.78	0.31	**6.19**	6.13	0.06
MetaCLIP L/14	ImageNet-trained	6.58	4.92	1.66	**12.74**	12.74	0.00
MetaCLIP L/14	CUB-trained	6.39	6.26	0.12	**18.43**	18.23	0.21
MetaCLIP L/14	NAB-trained	3.51	3.36	0.15	**6.92**	6.92	0.00
OpenAI L/14	ImageNet-trained	9.66	9.66	0.00	**19.88**	19.88	0.00
OpenAI L/14	CUB-trained	9.91	9.68	0.23	**29.65**	29.48	0.17
OpenAI L/14	NAB-trained	2.57	1.36	1.22	**14.18**	14.05	0.13
OpenAI B/32	ImageNet-trained	4.18	3.98	0.20	**8.98**	8.98	0.00
OpenAI B/32	CUB-trained	6.80	6.73	0.07	**11.86**	11.82	0.04
OpenAI B/32	NAB-trained	2.38	2.23	0.15	**3.80**	3.80	0.00

E.4 PAIRWISE CAPP TARGET RESULTS

CAPP is defined in §B.4.

Table S22: All six backbones (DFN/MetaCLIP backbones plus the additional OpenAI L/14 and B/32 backbones, marked below) fine-tuned separately on CUB-200 and on NABirds, each evaluated on its own target training ontology (CUB-trained → CUB-200, NAB-trained → NAB-555); PT is the pretrained checkpoint evaluated directly on the target dataset. Pairwise class-slot-macro accuracy at each method's own P1-selected deployment checkpoint; multi-way top-1 in Table S23. Structured supervision consistently improves controlled local attribute discrimination on the CUB target ontology, whereas NAB-target effects are more backbone-dependent.

Backbone	Train domain	PT	N	U	SSR	SSR+ND
DFN5B H/14-378	CUB-200	72.99	73.90	74.85	76.30	76.76
DFN5B H/14-378	NAB-555	65.92	67.56	67.52	65.65	65.79
DFN2B L/14	CUB-200	72.70	72.46	70.36	73.38	74.42
DFN2B L/14	NAB-555	64.23	63.35	64.08	63.67	64.67
MetaCLIP H/14	CUB-200	70.57	70.47	69.43	71.68	72.93
MetaCLIP H/14	NAB-555	63.14	60.88	63.48	65.63	66.76
MetaCLIP L/14	CUB-200	71.02	71.25	71.35	72.92	74.92
MetaCLIP L/14	NAB-555	63.57	59.89	63.06	65.74	66.76
Additional OpenAI backbones						
OpenAI L/14	CUB-200	66.63	67.51	69.11	70.86	75.05
OpenAI L/14	NAB-555	60.39	57.81	60.79	60.42	65.90

continued on next page

Backbone	Train domain	PT	N	U	SSR	SSR+ND
OpenAI B/32	CUB-200	64.22	65.32	67.19	68.06	69.84
OpenAI B/32	NAB-555	58.88	60.35	60.97	61.80	62.43

E.5 CAPP-MULTI TARGET RESULTS

Table S23: All six backbones (DFN/MetaCLIP backbones plus the additional OpenAI L/14 and B/32 backbones, marked below) fine-tuned separately on CUB-200 and on NABirds, each evaluated on its own target training ontology (CUB-trained → CUB-200, NAB-trained → NAB-555); PT is the pretrained checkpoint evaluated directly on the target dataset. Multi-way class-slot-macro top-1 accuracy at each method's own P1-selected deployment checkpoint; pairwise version in Table S22. Structured supervision consistently improves controlled local attribute discrimination on the CUB target ontology, whereas NAB-target effects are more backbone-dependent.

Backbone	Train domain	PT	N	U	SSR	SSR+ND
DFN5B H/14-378	CUB-200	49.13	49.44	50.15	52.97	53.53
DFN5B H/14-378	NAB-555	35.93	37.34	37.77	35.33	35.29
DFN2B L/14	CUB-200	46.32	46.24	43.19	47.37	49.70
DFN2B L/14	NAB-555	32.81	31.31	32.58	32.29	32.97
MetaCLIP H/14	CUB-200	45.26	44.85	41.64	46.19	47.89
MetaCLIP H/14	NAB-555	31.80	27.98	30.64	33.80	35.10
MetaCLIP L/14	CUB-200	44.69	44.35	44.22	47.85	50.44
MetaCLIP L/14	NAB-555	31.30	26.09	29.69	32.69	33.65
Additional OpenAI backbones						
OpenAI L/14	CUB-200	38.67	39.33	40.92	44.52	50.18
OpenAI L/14	NAB-555	28.45	26.52	29.16	29.05	34.12
OpenAI B/32	CUB-200	36.19	37.27	39.47	39.89	42.04
OpenAI B/32	NAB-555	28.89	28.84	29.62	29.68	29.83

E.6 CROSS-ONTOLOGY CAPP

Full and strict held-out populations are reported separately and never pooled (§B.5). Held-out CAPP effects are backbone- and transfer-direction-dependent.

Table S24: All six backbones (DFN/MetaCLIP backbones plus the additional OpenAI L/14 and B/32 backbones, marked below), evaluated cross-ontology in both directions: CUB-trained → NAB (full held-out NAB-555, strict dataset-class-disjoint NAB-348) and NAB-trained → CUB (full held-out CUB-200, strict dataset-class-disjoint CUB-59); "full" and "strict" are evaluation populations, not different training datasets, and are never pooled. PT is the pretrained checkpoint evaluated directly on each held-out population, and all methods are at their own P1-selected deployment checkpoint (pairwise class-slot-macro accuracy). Held-out CAPP effects are backbone- and transfer-direction-dependent; structured target-domain gains (Tables S22–S23) do not imply universal cross-ontology gains.

Backbone	Train→eval	Population	PT	N	U	SSR	SSR+ND
DFN5B H/14-378	CUB→NAB	NAB-555 (full)	65.95	67.23	68.05	67.66	66.56
DFN5B H/14-378	CUB→NAB	NAB-348 (strict)	65.63	67.05	68.22	66.85	65.92
DFN5B H/14-378	NAB→CUB	CUB-200 (full)	72.99	73.99	72.88	71.10	70.50
DFN5B H/14-378	NAB→CUB	CUB-59 (strict)	74.73	74.72	73.57	70.00	69.06

continued on next page

Backbone	Train→eval	Population	PT	N	U	SSR	SSR+ND
DFN2B L/14	CUB→NAB	NAB-555 (full)	64.24	64.07	63.56	64.01	63.97
DFN2B L/14	CUB→NAB	NAB-348 (strict)	63.70	63.23	62.86	63.44	63.09
DFN2B L/14	NAB→CUB	CUB-200 (full)	72.71	71.65	70.93	72.81	72.41
DFN2B L/14	NAB→CUB	CUB-59 (strict)	71.75	71.68	70.91	72.31	72.29
MetaCLIP H/14	CUB→NAB	NAB-555 (full)	63.13	63.27	63.29	64.70	64.72
MetaCLIP H/14	CUB→NAB	NAB-348 (strict)	62.60	61.98	62.62	64.29	64.12
MetaCLIP H/14	NAB→CUB	CUB-200 (full)	70.57	69.81	70.00	70.62	72.32
MetaCLIP H/14	NAB→CUB	CUB-59 (strict)	68.48	66.64	68.34	68.77	70.51
MetaCLIP L/14	CUB→NAB	NAB-555 (full)	63.61	63.26	63.36	64.95	64.69
MetaCLIP L/14	CUB→NAB	NAB-348 (strict)	63.04	62.47	62.45	64.25	63.92
MetaCLIP L/14	NAB→CUB	CUB-200 (full)	71.02	65.69	70.47	72.74	73.10
MetaCLIP L/14	NAB→CUB	CUB-59 (strict)	69.55	64.56	70.06	72.25	71.49
Additional OpenAI backbones							
OpenAI L/14	CUB→NAB	NAB-555 (full)	60.39	60.17	60.15	61.66	63.07
OpenAI L/14	CUB→NAB	NAB-348 (strict)	59.71	59.14	58.80	61.08	62.25
OpenAI L/14	NAB→CUB	CUB-200 (full)	66.63	64.80	68.18	68.17	71.46
OpenAI L/14	NAB→CUB	CUB-59 (strict)	66.67	63.46	68.44	68.74	72.35
OpenAI B/32	CUB→NAB	NAB-555 (full)	58.91	60.14	59.93	60.55	61.29
OpenAI B/32	CUB→NAB	NAB-348 (strict)	58.40	59.56	58.80	60.40	60.74
OpenAI B/32	NAB→CUB	CUB-200 (full)	64.22	65.50	65.56	67.77	68.42
OpenAI B/32	NAB→CUB	CUB-59 (strict)	64.71	66.01	62.72	65.34	67.35

E.7 DIRECT NAME-ACCESS ABLATION: QUANTITATIVE RESULTS

The FULL/PARTIAL/NO variants are defined in §A.5. They control ATTRIBUTE-route NAME access only; EOT remains globally name-accessible, so even NO is not globally name-free.

Direct name-access ablation results. Tables S25–S28 give the FULL/PARTIAL/NO results (single runs, DFN5B H/14-378, CUB-trained). Restricting direct NAME access from the ATTRIBUTE route increases target P3 and CAPP and makes the attribute route more similar to the pretrained name-free profile geometry and less similar to class-name geometry, while ordinary P2 still collapses under name-swap in every variant. NO is not globally name-free: EOT remains globally name-accessible. The name-access effect of this ablation and the late-specialization analyses in §G are mechanistically distinct and should not be conflated. Caveats: PARTIAL deploys at epoch 75 versus epoch 20 for the others; NAB differences of about 1 pp should not be over-read.

Variant	Eval.	P3 PT	P3 peak (ep.)	P3 deploy (ep.)	Gain_{peak}	Gain_{deploy}	Forget
FULL	CUB-200	34.36	36.16 (4)	34.17 (20)	1.79	-0.19	1.99
	NAB-348	13.81	16.26 (12)	15.91 (20)	2.45	2.10	0.35
PARTIAL	CUB-200	34.45	37.80 (12)	36.97 (75)	3.35	2.52	0.83
	NAB-348	13.87	17.01 (12)	13.38 (75)	3.14	-0.49	3.63
NO	CUB-200	34.35	40.42 (20)	40.42 (20)	6.08	6.08	0.00
	NAB-348	13.85	17.32 (12)	16.90 (20)	3.48	3.06	0.42

Table S25: Direct name-access variants (FULL / PARTIAL / NO), DFN5B H/14-378 fine-tuned on CUB-200, one run per variant (seed 0). Only the routing mask differs. FULL: the ATTRIBUTE route sees the full NAME segment; PARTIAL: only the NAME_END boundary signal; NO: no NAME access (identical to the canonical SSR run). EOT remains globally name-accessible in all three. Deployment is the CUB-P1-selected epoch (FULL 20, PARTIAL 75, NO 20), so PARTIAL's deployment epoch differs from the others.

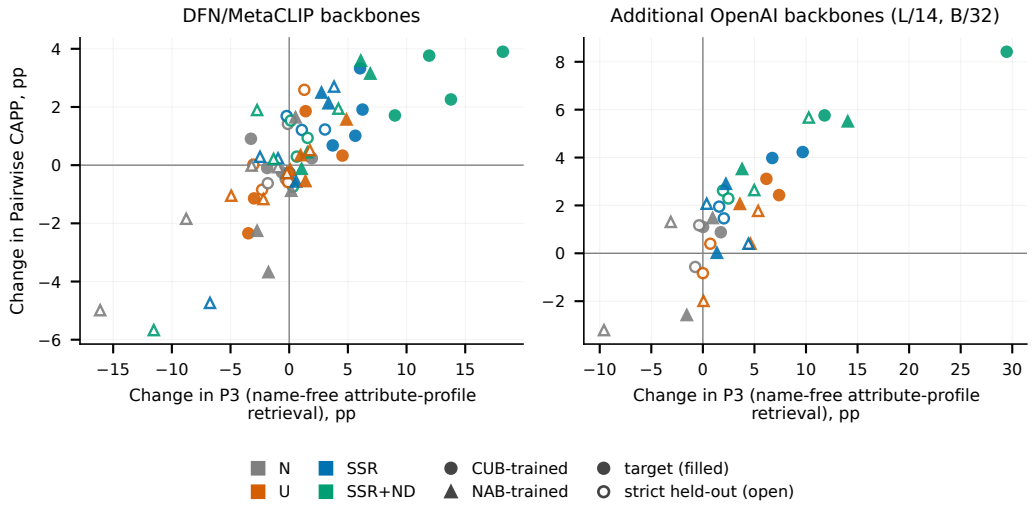


Figure S3: **Name-free attribute-profile retrieval (P3) change versus Pairwise CAPP change** across intervention conditions, each relative to the pretrained model, at every run's own P1-selected deployment checkpoint. Left: DFN/MetaCLIP backbones; right: additional OpenAI L/14 and B/32 backbones (never pooled). Colors: N, U, SSR, SSR+ND; circles CUB-trained and triangles NABirds-trained; filled markers are the training-domain target and open markers the strict dataset-class-disjoint held-out set (NAB-348 for CUB-trained, CUB-59 for NABirds-trained; full held-out sets are not plotted). Across many intervention conditions, gains in name-free attribute-profile retrieval and local counterfactual discrimination tend to move in the same direction, although the probes measure distinct capabilities. No correlation coefficient is computed. Delta values use the pretrained result from the same evaluation job as each point and may therefore differ slightly from differences reconstructed from separately tabulated evaluations.

Variant	Eval.	Ep. 0 (PT)	Ep. 8	Ep. 20	Ep. 100
FULL	CUB-200	72.99	74.28	74.85	73.74
	NAB-348	65.62	65.65	66.72	65.81
PARTIAL	CUB-200	72.97	75.14	75.59	74.32
	NAB-348	65.62	65.94	67.05	66.07
NO	CUB-200	72.99	75.44	76.31	75.06
	NAB-348	65.62	65.94	66.85	65.60

Table S26: Pairwise CAPP (class-slot-macro accuracy, %) for the name-access variants at matched epochs. On the CUB target, CAPP follows NO > PARTIAL > FULL at epochs 8, 20, and 100 (differences of 0.3–1.5 pp); on NAB-348 the variants are nearly indistinguishable and the gain is transient.

Variant	Eval.	Dep. ep.	P2 ord.	P2 name-sw.	P2 attr-sw.	Name-sw. damage	Attr-sw. damage
FULL	CUB-200	20	89.90	0.05	86.52	89.85	3.38
	NAB-348	20	79.66	0.30	69.82	79.35	9.83
PARTIAL	CUB-200	75	90.28	0.07	88.64	90.21	1.64
	NAB-348	75	73.42	0.23	65.60	73.19	7.83
NO	CUB-200	20	89.51	0.12	85.42	89.39	4.09
	NAB-348	20	78.97	0.33	68.39	78.64	10.59

Table S27: Ordinary, name-swapped, and attribute-swapped P2 (R@1 %) at each variant's deployment epoch. Name-swap collapses P2 to $\leq$0.33% in every variant, so restricting the ATTRIBUTE route's name access does not remove the class-name dominance of ordinary P2.

Variant	Attr. end-mean	Name end-mean	Attr. pool-proj	Name pool-proj
FULL	0.39	0.84	0.40	0.83
PARTIAL	0.40	0.76	0.40	0.76
NO	0.65	0.35	0.68	0.36

Table S28: Representational similarity (RSA) of the ATTRIBUTE-route representation at the last layer, epoch 20, to the pretrained name-free attribute-profile geometry and to class-name geometry, under two readouts (att_end_mean and att_pool_proj). Restricting direct NAME access raises similarity to the attribute-profile geometry and lowers similarity to class-name geometry under both readouts (main-text summary values use att_end_mean for the attribute profile and att_pool_proj for class name; both are shown here).

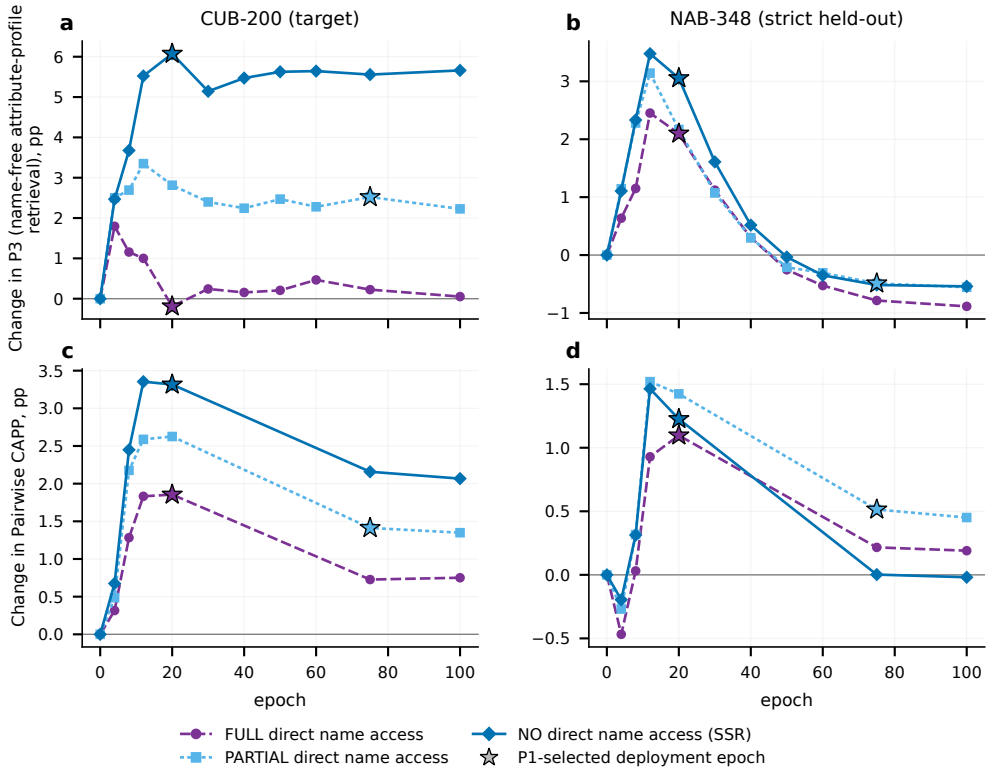


Figure S4: **Direct name-access ablation trajectories,** DFN5B H/14-378 fine-tuned on CUB-200 (one run per variant, single seed). Rows: P3 (name-free attribute-profile retrieval) and Pairwise CAPP; columns: CUB-200 (target) and strict held-out NAB-348. FULL, PARTIAL and NO direct name access denote how much of the NAME segment the ATTRIBUTE routes may read (Appendix A.5); NO is the canonical SSR run. Values are the change from each run's own pretrained epoch-0 evaluation; stars mark each run's P1-selected deployment epoch (FULL 20, PARTIAL 75, NO 20). P3 is evaluated at 11 epochs and CAPP at 7 (0, 4, 8, 12, 20, 75, 100), so the CAPP segment between epochs 20 and 75 is a straight line across unmeasured epochs. Restricting direct name access from the ATTRIBUTE route changes P3 and CAPP acquisition and trajectory behavior, supporting a role for representation and routing rather than content alone; NO is not name-free globally, because the EOT summary still reads NAME_END, and PARTIAL deploys later, so deployment-level comparisons are timing-confounded.

F Capability-Specific Timing and Specialization

F.1 P1 and P3 trajectory atlas

Figures S5–S9 show raw P1 (class-name retrieval) and P3 (name-free attribute-profile retrieval) trajectories in matched rows, so that acquisition, capability-specific peak timing, post-peak decline, the divergence between P1 and P3, and the position of the deployment checkpoint can be read directly. All values are absolute image→text R@1 (%) at the saved checkpoints only, connected by straight lines without interpolation; the dashed line is the pretrained model; the large outlined marker on each curve is that run's existing P1-selected deployment checkpoint (target-domain P1; no separate P1 optimum is selected, and curves are not normalized). Every panel is a single run per method and backbone, and the backbones are heterogeneous models, not replicates. P2 trajectories are omitted here because P2 has a different role in the paper. Not every trajectory declines: for OpenAI L/14 trained on ImageNet the deployment checkpoint is the final evaluated milestone and coincides with the P3 maximum for all four conditions on the target and same-ontology evaluations (§F.6; those populations are not plotted here).

F.2 Representative P3 / CAPP / deployment timing

Table S29 gives the P3 peak epoch, Pairwise-CAPP peak epoch, and P1-selected deployment epoch for the DFN5B H/14-378 CUB-trained runs (plotted in Figure S10) visualizes these trajectories. This CAPP-timing observation is a representative single-run result; CAPP peak timing was not evaluated across six backbones. Values are consistent with the reported trajectory results. The SSR+ND run has 8 dense checkpoints versus 11 for the other methods, so its peak-epoch resolution is coarser.

F.3 Six-backbone before/at/after deployment counts

Peak and deployment milestones are compared *categorically* (before / at / after) within each run's own schedule; absolute epochs or milestone indices are never pooled across incompatible schedules. Summed over all six backbones and four methods, the counts (before/at/after) are 11/13/0 for the ImageNet target (24 cells), 40/32/0 for the same-ontology shifts (72 cells), and 65/7/0 for the cross-ontology evaluations (72 cells). Target and same-ontology optima occur at or after deployment more often, whereas cross-ontology peaks occur before deployment in the large majority of cells.

F.4 Prevalence and magnitude of early P3 optima

Table S30 gives the full before/at/after timing counts by population and method (target 11/13/0 of 24, same-ontology 40/32/0 of 72, cross-ontology 65/7/0 of 72), which underlie Figure 4b; only target and cross-ontology are plotted there. Table S31 gives the peak-to-deployment P3 decrease among cells whose peak precedes deployment (Figure 4c shows the cross-ontology cells), including the > 0.5 pp counts, the target summaries that are too small to plot, and cross-ontology medians by dataset (CUB-200 1.78 pp, NAB-348 1.76 pp, FineView Butterfly 0.58 pp).

F.5 Per-backbone timing reconstruction

Table S32 gives the per-backbone counts whose column sums reproduce the population totals (Table S30) behind Figure 4b. Per-cell peak and deployment milestones are in Tables S11–S13.

F.6 OpenAI L/14 evaluated-trajectory boundary case

For OpenAI L/14 the P1-selected deployment checkpoint is the final evaluated ImageNet milestone (11), which also coincides with the P3 maximum in all 16 target and same-ontology cells (Table S33); hence no post-peak decline was observed within the evaluated trajectory. This describes the audited schedule and does not establish that specialization cannot occur for this backbone.

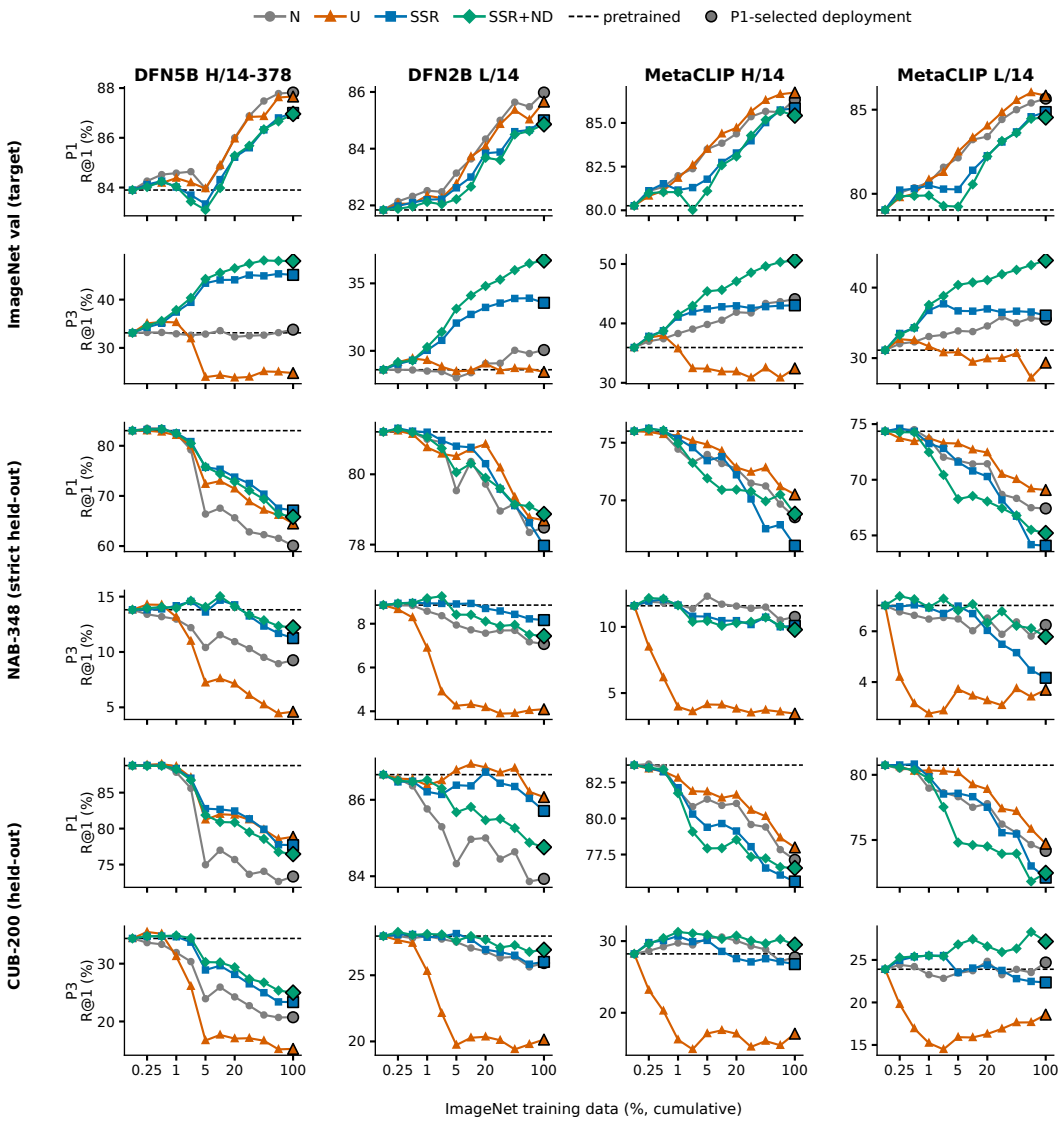


Figure S5: **ImageNet-trained P1/P3 trajectories, DFN/MetaCLIP backbones** (columns: DFN5B H/14-378, DFN2B L/14, MetaCLIP H/14, MetaCLIP L/14). For each evaluation population, the top row is P1 and the bottom row P3: ImageNet val (target), NAB-348 (strict dataset-class-disjoint held-out) and CUB-200 (held-out). The horizontal axis is the cumulative ImageNet training-data fraction at the 11 geometric milestones (not epochs). Deployment markers show the ImageNet-val-P1-selected checkpoint, which is shared by every population of a run; P3 uses the templates in Appendix B.3 (ImageNet: `a object with {attributes}`; bird evaluations: `a bird with {attributes}`).

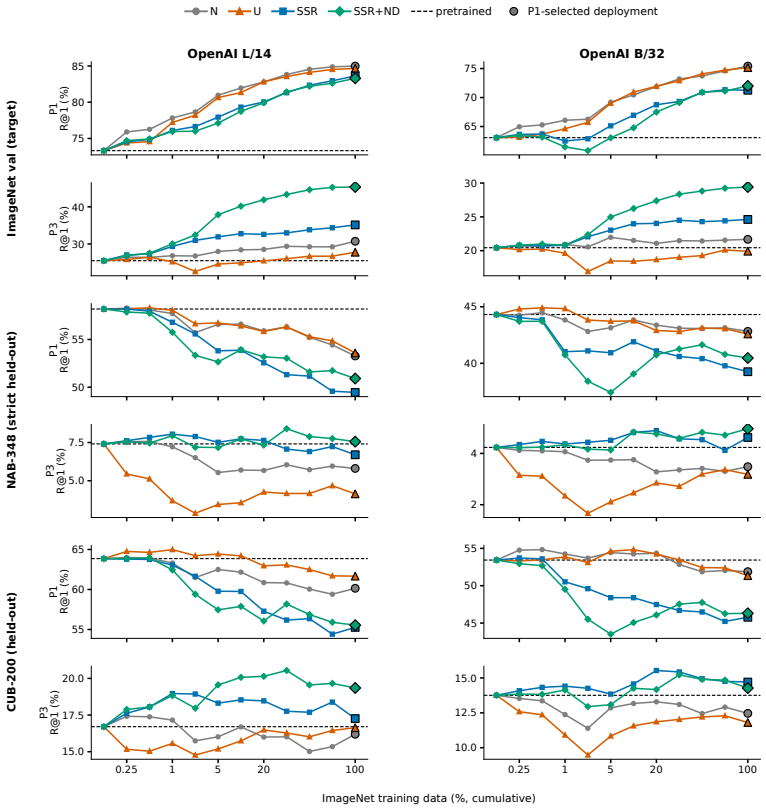


Figure S6: **ImageNet-trained P1/P3 trajectories, additional OpenAI backbones** (OpenAI L/14 and B/32; the two families are never pooled). Layout, populations and definitions as in Figure S5. The OpenAI deployment checkpoint is the ImageNet-val-P1 argmax over the 11 milestones; for L/14 it is the final milestone.

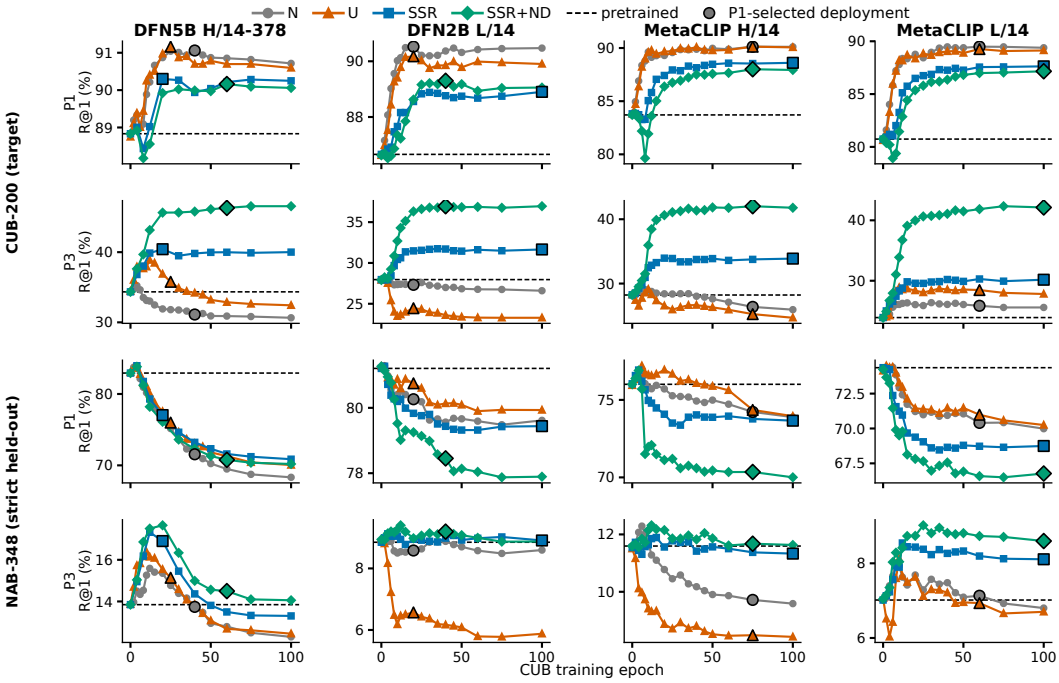


Figure S7: **CUB-trained P1/P3 trajectories, DFN/MetaCLIP backbones.** Rows: CUB-200 (target) and NAB-348 (strict dataset-class-disjoint held-out), each with P1 above P3. Horizontal axis: CUB training epoch. Deployment markers show each run's CUB-P1-selected checkpoint. The DFN5B SSR+ND curve is the canonical SSR+ND run; dense fine-grained trajectories for the DFN/MetaCLIP backbones come from the per-epoch trajectory evaluations rather than a consolidated table, and the full NAB-555 held-out population has no P3 trajectory for the DFN/MetaCLIP backbones (it is available for OpenAI in Figure S9).

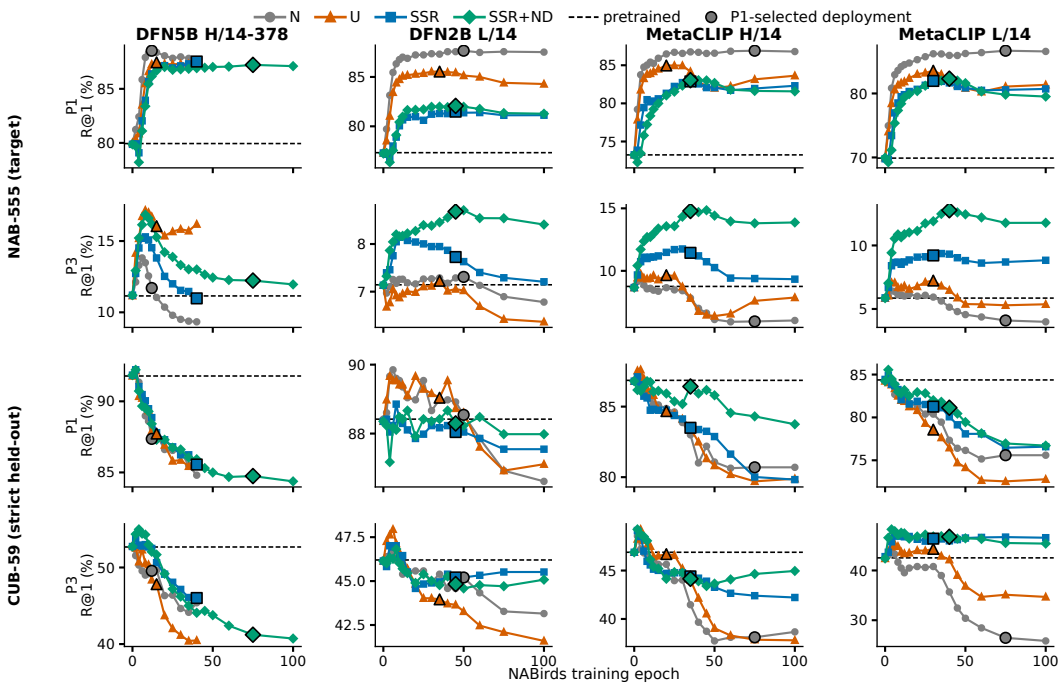


Figure S8: **NABirds-trained P1/P3 trajectories, DFN/MetaCLIP backbones.** Rows: NAB-555 (target) and CUB-59 (strict dataset-class-disjoint held-out), each with P1 above P3. Horizontal axis: NABirds training epoch (DFN5B N/U/SSR were evaluated to epoch 40). Deployment markers show each run's NAB-P1-selected checkpoint. Definitions and single-run caveats as in Figure S7.

Method	Eval.	P3 peak (ep.)	P3 deploy (ep.)	CAPP peak (ep.)	CAPP deploy (ep.)
N	CUB-200 (target)	4	40	15	40
N	NAB-348 (held-out)	12	40	20	40
U	CUB-200 (target)	12	25	15	25
U	NAB-348 (held-out)	10	25	15	25
SSR	CUB-200 (target)	20	20	12	20
SSR	NAB-348 (held-out)	12	20	12	20
SSR+ND	CUB-200 (target)	75	60	20	60
SSR+ND	NAB-348 (held-out)	20	60	12	60

Table S29: Dense P3 and Pairwise-CAPP peak/deployment epochs, DFN5B H/14-378 fine-tuned on CUB-200, evaluated on CUB-200 (target) and NAB-348 (strict held-out); underlies Figure S10. Deployment epoch is each method's own P1-selected checkpoint, shared by construction between the P3 and CAPP columns. CAPP is Pairwise class-slot-macro accuracy. Single run per method.

Table S32: Per-backbone counts of evaluation cells (over the four methods N/U/SSR/SSR+ND and the datasets in each population) whose P3 peak milestone falls before, at, or after the run's P1-selected deployment milestone. T = ImageNet val target (4 cells per backbone), S = same-ontology shifts ImageNet-V2/-R/-Sketch (12 cells), X = cross-ontology CUB-200/NAB-348/FVB (12 cells). Classification compares peak and deployment milestone indices from Tables S11–S13; milestones are cumulative-data indices, not epochs, and are never pooled as absolute values across schedules. Column sums over backbones give the population totals in Table S30.

Backbone	T-bef	T-at	T-aft	S-bef	S-at	S-aft	X-bef	X-at	X-aft
DFN5B H/14-378	3	1	0	10	2	0	12	0	0

continued on next page

Backbone	T-bef	T-at	T-aft	S-bef	S-at	S-aft	X-bef	X-at	X-aft
DFN2B L/14	2	2	0	6	6	0	9	3	0
MetaCLIP H/14	1	3	0	6	6	0	11	1	0
MetaCLIP L/14	3	1	0	7	5	0	11	1	0
OpenAI L/14	0	4	0	0	12	0	11	1	0
OpenAI B/32	2	2	0	11	1	0	11	1	0
All six	11	13	0	40	32	0	65	7	0

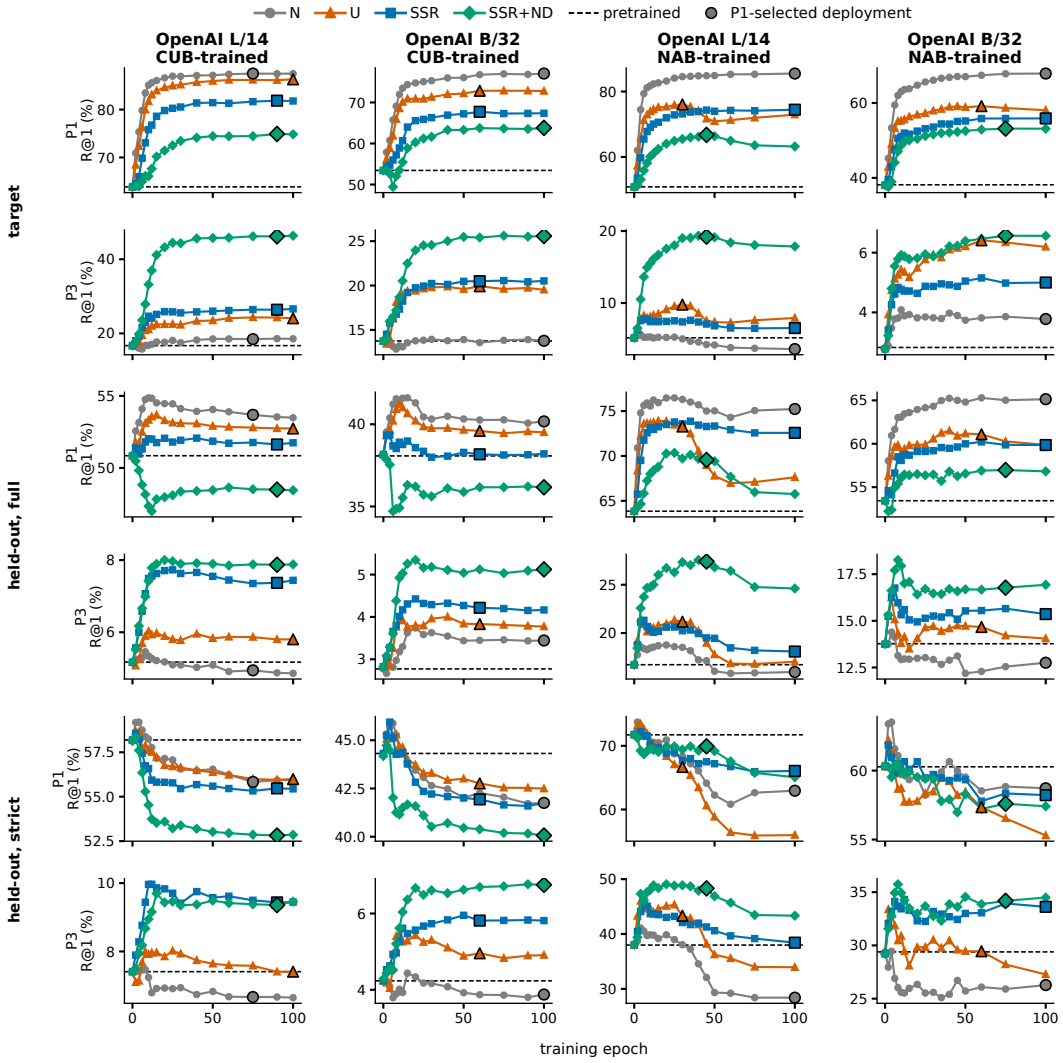


Figure S9: **OpenAI fine-grained P1/P3 trajectories** (columns: OpenAI L/14 and B/32, CUB-trained and NABirds-trained). Rows: the training-domain target (CUB-200 or NAB-555), the full held-out population (NAB-555 or CUB-200; not dataset-class-disjoint) and the strict held-out population (NAB-348 or CUB-59; dataset-class-disjoint), each with P1 above P3; full and strict populations are never pooled. Deployment markers show each run's target-domain-P1-selected epoch (argmax over the saved epochs after epoch 0). All four methods are single runs; OpenAI CUB-trained runs use weight decay 0.2 for all four methods.

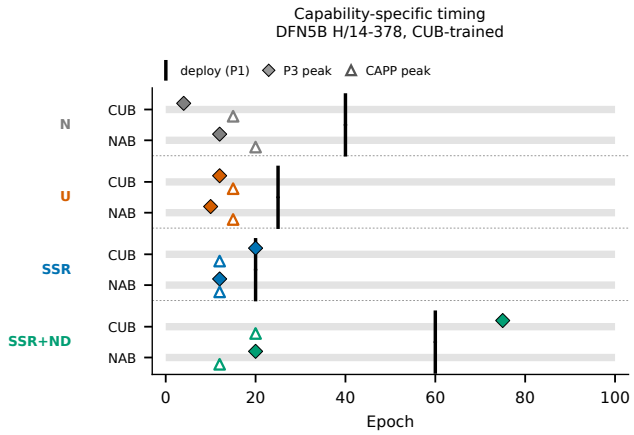


Figure S10: P3 peak (diamond), Pairwise-CAPP peak (open triangle), and P1-selected deployment epoch (tick) for each method and evaluation set, DFN5B H/14-378 fine-tuned on CUB-200 (single run per method). The three need not coincide.

Population	Method	Cells	Before	At	After
target	N	6	2	4	0
	U	6	5	1	0
	SSR	6	3	3	0
	SSR+ND	6	1	5	0
	all	24	11	13	0
same-ontology	N	18	6	12	0
	U	18	15	3	0
	SSR	18	13	5	0
	SSR+ND	18	6	12	0
	all	72	40	32	0
cross-ontology	N	18	16	2	0
	U	18	17	1	0
	SSR	18	17	1	0
	SSR+ND	18	15	3	0
	all	72	65	7	0

Table S30: Timing of each evaluation distribution's own P3-peak milestone relative to the run's P1-selected deployment milestone, by method (six ImageNet-trained backbones). Target = ImageNet val (6 cells per method); same-ontology = ImageNet-V2/-R/-Sketch (18); cross-ontology = CUB-200, NAB-348, FineView Butterfly (18). Milestones are cumulative-data indices and are compared only categorically. Panel b of Figure 4 plots "before" for target and cross-ontology. Six heterogeneous backbones, not replicates.

Population	Method	Valid n	Median (pp)	IQR	> 0.5 pp
target	N	2	0.34	[0.32, 0.36]	0
	U	5	3.32	[1.00, 5.56]	4
	SSR	3	0.34	[0.28, 1.00]	1
	SSR+ND	1	0.12	[0.12, 0.12]	0
	all	11	0.38	[0.31, 2.49]	5
cross-ontology	N	16	1.40	[0.62, 2.01]	14
	U	17	1.21	[0.58, 5.04]	13
	SSR	17	1.71	[0.83, 2.87]	15
	SSR+ND	15	1.35	[0.90, 1.80]	12
	all	65	1.35	[0.64, 2.82]	54
by dataset	CUB-200	23	1.78	[1.16, 5.00]	22
	NAB-348	23	1.76	[0.82, 2.85]	20
	FineView Butterfly	19	0.58	[0.47, 1.38]	12

Table S31: Peak-to-deployment P3 decrease, $\mathrm{P3}_{\mathrm{peak}} - \mathrm{P3}_{\mathrm{deploy}}$ (absolute pp), over cells whose own P3 peak precedes deployment ($t_{P3} < t_{\mathrm{deploy}}$); cells with peak at or after deployment have no evaluated post-peak interval and are excluded. Only 11 target cells qualify (n = 1–5 per method), too few for stable per-method target summaries, so Figure 4c shows cross-ontology cells only. The > 0.5 pp count is a descriptive robustness count, not a significance test. The by-dataset rows are cross-ontology cells pooled over methods. OpenAI L/14 target and same-ontology cells all have the peak at the final milestone (Table S33), so absence of observed decline there is bounded by the evaluated schedule.

Table S33: OpenAI L/14 (ImageNet-trained): peak and deployment milestone indices for the ImageNet target and same-ontology shifts. In every one of these 16 cells the P1-selected deployment checkpoint is the final evaluated milestone (11) and coincides with the P3 maximum, so no post-peak interval exists in the evaluated trajectory; this is a statement about the audited schedule, not evidence that specialization cannot occur for this backbone. Cross-ontology cells (Table S13) do show peaks before deployment.

Dataset	Method	Peak ms.	Deploy ms.	Gain_{peak}	Forget
ImageNet val	N	11	11	5.24	0.00
	U	11	11	2.28	0.00
	SSR	11	11	9.66	0.00
	SSR+ND	11	11	19.88	0.00
ImageNet-V2	N	11	11	4.74	0.00
	U	11	11	1.89	0.00
	SSR	11	11	8.75	0.00
	SSR+ND	11	11	16.95	0.00
ImageNet-R	N	11	11	6.47	0.00
	U	11	11	5.83	0.00
	SSR	11	11	10.88	0.00
	SSR+ND	11	11	21.63	0.00
ImageNet-Sketch	N	11	11	5.55	0.00
	U	11	11	2.46	0.00
	SSR	11	11	7.31	0.00
	SSR+ND	11	11	12.47	0.00

G Analyses of Late Specialization

Except for §G.1 (broad trajectory grid) and §G.5 (additional swaps on other runs), all results in this section are single-run observations for the representative DFN5B H/14-378 SSR+ND model fine-

tuned on CUB-200 and evaluated on NAB-348 (and CUB-200 where stated), plus an independent text-frozen U run where stated. They are not claimed to generalize across backbones or training domains, and no replication is implied. The preferred interpretation is reduced cross-modal semantic accessibility; the analyses do not show that semantic information is erased.

G.1 Broad direct-movement analysis (N/U/SSR grid)

This subsection supports Figure 5a and, unlike the other subsections, covers the broad N/U/SSR trajectory grid (4 backbones $\times$ 3 training domains $\times$ 3 methods; 120 train/eval/backbone/method cells, 84 held-out; SSR+ND is not included). For each evaluation distribution the P3-peak checkpoint is that distribution's own P3 maximum (epoch 0 excluded) and the deployment checkpoint is the run's P1-selected checkpoint; no checkpoint was substituted. Table S34 gives the temporal order of the two checkpoints; the primary late-specialization analysis uses only $t_{P3} < t_{\text{deploy}}$ cells.

Movement definitions. For each transformer block l of a encoder, $D_{\text{pre}}(l) = \|W_{\text{peak}}(l) - W_{\text{PT}}(l)\|_F / \|W_{\text{PT}}(l)\|_F$ and $D_{\text{post}}(l) = \|W_{\text{deploy}}(l) - W_{\text{peak}}(l)\|_F / \|W_{\text{PT}}(l)\|_F$, averaged over blocks; both use the pretrained norm so the two phases are directly comparable, and both are net checkpoint displacements, not SGD path lengths. Image movement uses `visual.transformer.resblocks.*` (32 blocks, 629.7M parameters for DFN5B and MetaCLIP H/14; 24 blocks, 302.3M for DFN2B and MetaCLIP L/14). Text movement uses the shared pretrained text transformer `transformer.resblocks.*` (24 blocks, 302.3M parameters; 12 blocks, 85.1M). Excluded from the text metric because they have no pretrained counterpart or are not comparable across N/U/SSR: AIP and StructuredTextFusion parameters, the 4–6 newly appended SSR token-embedding rows, positional embeddings, and the logit scale. The main-text quantity is the joint pretrained-backbone movement $D_{\text{joint}} = (D_{\text{image}} + D_{\text{text}})/2$ (not total model movement).

Results. Pre-peak movement has no consistent relationship with acquisition for either encoder: the sign differs by training domain (image: pooled $r = -0.17$; CUB $+0.49$, NABirds -0.72, ImageNet $+0.33$; text: pooled -0.14; CUB $+0.63$, NABirds -0.73, ImageNet $+0.35$; joint pooled -0.16). Post-peak movement is moderately and positively associated with subsequent P3 loss (image $r = 0.43$, $\rho = 0.51$; text $r = 0.54$, $\rho = 0.64$; joint $r = 0.50$, $\rho = 0.61$; $n = 78$; Table S35). Image and text movement are strongly coupled (post-peak $r = 0.85$, $\rho = 0.89$; within-backbone standardized $r = 0.95$; pre-peak $r = 0.95$), and their apparent relative importance depends on normalization (Table S36): pooled and domain-standardized regressions favor text, while within-backbone standardization gives near-equal image and text associations with variance-inflation factors up to about 11. Broad encoder-specific attribution is therefore not identifiable, and image-side localization is supported only by the representative encoder-state intervention below. Robustness: the text transformer plus projection and final LayerNorm gives $r = 0.54$ and the pretrained token-embedding rows alone $r = 0.41$.

An earlier radial quantity (distance-to-PT at deployment minus distance-to-PT at the peak, image encoder) correlates with the direct image post-peak quantity at $r = 0.95$ over the 107 peak-before-deployment cells, with the direct value a median $1.11\times$ larger, and gave a somewhat stronger held-out association with the post-peak decrease ($r \approx 0.56$ versus 0.43); the direct displacement is used throughout. Normalizing by the peak norm instead of the pretrained norm gives the same correlation to three decimals, because checkpoint changes are below 1% of the parameter norm. The ImageNet-trained image-encoder association is near zero when pooled partly because of between-backbone scale structure (within individual ImageNet backbones r is about $+0.26$ to $+0.49$). All statistics are descriptive associations across heterogeneous conditions, not causal or replicate-based estimates.

G.2 Held-out P1 over the same P3-peak→deployment interval

As a robustness check on Figure 5a, we asked whether the same joint backbone movement also tracks held-out P1 over the identical interval. For each of the 78 held-out trajectory cells with $t_{P3} < t_{\text{deploy}}$, P1 is evaluated at that cell's P3-peak checkpoint and at its P1-selected deployment checkpoint, and *P1 decrease* is the P1 value at the P3-peak checkpoint minus the value at deployment (positive = held-out P1 got worse); no separate P1-optimal checkpoint is selected. The joint movement is the same quantity as in Figure 5a. Movement is more strongly associated with the P1 decrease than with the P3 decrease over the same cells (Figure S11: Pearson $r = 0.73$, Spear-

Training domain	Cells	$t_{P3} < t_{\text{deploy}}$	$t_{P3} = t_{\text{deploy}}$	$t_{P3} > t_{\text{deploy}}$
CUB-trained	held-out	23	0	1
CUB-trained	target	11	1	0
ImageNet-trained	held-out	33	3	0
ImageNet-trained	target	10	2	0
NABirds-trained	held-out	22	1	1
NABirds-trained	target	8	3	1
All training domains	held-out	78	4	2
All training domains	target	29	6	1

Table S34: Temporal order of each evaluation distribution's own P3-peak checkpoint and the P1-selected deployment checkpoint over the 120 N/U/SSR cells (4 backbones × 3 training domains × 3 methods × evaluation sets). Only $t_{P3} < t_{\text{deploy}}$ cells are temporally valid "post-peak" trajectories.

	Pre-peak vs. $\text{Gain}_{\text{peak}}$				Post-peak vs. $\text{Forget}_{\text{postpeak}}$			
Subset	n	Image	Text	Joint	n	Image	Text	Joint
Held-out pooled (primary)	84	−0.17/+0.33	−0.14/+0.35	−0.16/+0.32	78	+0.43/+0.51	+0.54/+0.64	+0.50/+0.61
CUB-trained	24	+0.49/+0.58	+0.63/+0.69	+0.58/+0.63	23	+0.42/+0.49	+0.68/+0.72	+0.58/+0.59
NABirds-trained	24	−0.72/−0.65	−0.73/−0.70	−0.73/−0.72	22	+0.64/+0.64	+0.56/+0.79	+0.64/+0.75
ImageNet-trained	36	+0.33/+0.57	+0.35/+0.57	+0.34/+0.53	33	+0.10/−0.02	+0.56/+0.54	+0.39/+0.43
N	28	−0.25/+0.31	−0.23/+0.35	−0.24/+0.32	26	+0.45/+0.54	+0.40/+0.58	+0.44/+0.60
U	28	+0.01/+0.49	+0.08/+0.52	+0.04/+0.48	26	+0.33/+0.16	+0.61/+0.46	+0.48/+0.35
SSR	28	−0.31/−0.01	−0.31/−0.06	−0.31/−0.09	26	+0.59/+0.80	+0.71/+0.85	+0.66/+0.83
DFN5B H/14-378	21	−0.29/+0.32	−0.29/+0.25	−0.29/+0.25	21	+0.65/+0.83	+0.64/+0.83	+0.65/+0.83
DFN2B L/14	21	−0.08/+0.68	−0.04/+0.64	−0.06/+0.66	17	+0.51/+0.63	+0.55/+0.69	+0.55/+0.69
MetaCLIP H/14	21	−0.25/+0.09	−0.22/+0.17	−0.24/+0.08	21	+0.41/+0.44	+0.39/+0.43	+0.40/+0.44
MetaCLIP L/14	21	−0.15/+0.12	−0.12/+0.07	−0.13/+0.09	19	+0.29/+0.32	+0.21/+0.32	+0.26/+0.32
Text-component robustness (pooled held-out; Pearson/Spearman)								
Text transformer only	84		−0.14/+0.35		78		+0.54/+0.64	
Text transformer + projection + final LayerNorm	84		−0.12/+0.40		78		+0.54/+0.61	
Pretrained token-embedding rows only	84		−0.15/+0.27		78		+0.41/+0.53	

Table S35: Encoder-separated and joint movement analysis (Pearson r / Spearman ρ), held-out cells of the N/U/SSR grid; post-peak columns use $t_{P3} < t_{\text{deploy}}$ cells. Image movement: mean over image-encoder blocks (`visual.transformer.resblocks.*`) of $\|\Delta W\|_F / \|W_{\text{PT}}\|_F$. Text movement: the same quantity over the shared pretrained text transformer (`transformer.resblocks.*`; 24 blocks for DFN5B/MetaCLIP H/14, 12 for DFN2B/MetaCLIP L/14). Joint = (image + text)/2, the "pretrained-backbone movement" used in Figure 5a. Excluded because they lack a pretrained analogue or are not comparable: AIP and StructuredTextFusion parameters, the newly appended SSR token-embedding rows (4–6), positional embeddings, and the logit scale. Descriptive associations across heterogeneous backbones; not replicates.

		Pearson r			Partial r		Std. coef.			
Subset / normalization	n	Img–F	Txt–F	Img–Txt	Img|Txt	Txt|Img	Img	Txt	R^2	VIF
Pooled (raw)	78	+0.43	+0.54	+0.85	−0.06	+0.36	−0.10	+0.62	0.29	3.5
CUB-trained	23	+0.42	+0.68	+0.72	−0.14	+0.60	−0.15	+0.79	0.48	2.1
NABirds-trained	22	+0.64	+0.56	+0.76	+0.40	+0.14	+0.51	+0.17	0.42	2.4
ImageNet-trained	33	+0.10	+0.56	+0.45	−0.22	+0.59	−0.20	+0.66	0.35	1.3
Domain-standardized	78	+0.34	+0.60	+0.62	−0.04	+0.52	−0.04	+0.62	0.36	1.6
Within-backbone standardized	78	+0.47	+0.45	+0.95	+0.16	+0.00	+0.46	+0.01	0.22	10.8
Within backbone × domain standardized	78	+0.54	+0.56	+0.90	+0.10	+0.21	+0.19	+0.39	0.33	5.1

Table S36: Post-peak image versus text movement, held-out $t_{P3} < t_{\text{deploy}}$ cells (n = 78 pooled). F = $\text{Forget}_{\text{postpeak}}$; Img–Txt is the image/text movement correlation; Partial r controls the other encoder; Std. coef. are from the descriptive regression $\text{F} \sim z(D_{\text{image}}) + z(D_{\text{text}})$. Image and text movement are strongly coupled, and the apparent relative importance of the encoders depends on normalization (text dominates pooled and domain-standardized, image and text are indistinguishable within backbone; VIF up to about 11), so broad encoder attribution is not identifiable. Descriptive only; not a population-level regression or causal claim.

	Pre-peak: D_{pre} vs. $\mathrm{Gain}_{\mathrm{peak}}$			Post-peak: D_{post} vs. $\mathrm{Forget}_{\mathrm{postpeak}}$		
Subset	n	r	ρ	n	r	ρ
Held-out, all (primary)	84	−0.17	+0.33	78	+0.43	+0.51
All valid cells	108	−0.15	+0.28	107	+0.44	+0.52
CUB-trained (held-out)	24	+0.49	+0.58	23	+0.42	+0.49
NABirds-trained (held-out)	24	−0.72	−0.65	22	+0.64	+0.64
ImageNet-trained (held-out)	36	+0.33	+0.57	33	+0.10	−0.02
CUB+NABirds-trained (held-out)	48	−0.37	−0.08	45	+0.65	+0.69
N (held-out)	28	−0.25	+0.31	26	+0.45	+0.54
U (held-out)	28	+0.01	+0.49	26	+0.33	+0.16
SSR (held-out)	28	−0.31	−0.01	26	+0.59	+0.80
DFN5B H/14-378 (held-out)	21	−0.29	+0.32	21	+0.65	+0.83
DFN2B L/14 (held-out)	21	−0.08	+0.68	17	+0.51	+0.63
MetaCLIP H/14 (held-out)	21	−0.25	+0.09	21	+0.41	+0.44
MetaCLIP L/14 (held-out)	21	−0.15	+0.12	19	+0.29	+0.32
Robustness (post-peak, held-out, $t_{P3} < t_{\mathrm{deploy}}$)						
Peak-normalized D_{post}				78	+0.43	+0.51
Distance-to-pretrained radial quantity (distance-to-PT difference)				78	+0.56	+0.63
Distance-to-pretrained radial, layer-aggregated				78	+0.50	+0.57
Full grid, symmetric distance (held-out, incl. $t_{P3} \geq t_{\mathrm{deploy}}$)				84	+0.46	+0.60
Domain-demeaned pooled				78	+0.31	+0.26

Table S37: Direct image-encoder movement versus P3 change across the N/U/SSR trajectory grid (Pearson r, Spearman ρ; absolute percentage points). D_{pre} and D_{post} are the layer-averaged Frobenius displacements $\|W_{\mathrm{peak}} - W_{\mathrm{PT}}\|/\|W_{\mathrm{PT}}\|$ and $\|W_{\mathrm{deploy}} - W_{\mathrm{peak}}\|/\|W_{\mathrm{PT}}\|$ over all image-encoder transformer-block parameters (attention, MLP, LayerNorm; projection, embeddings, text encoder and SSR modules excluded); net checkpoint displacement, not SGD path length. Post-peak rows use held-out cells with $t_{P3} < t_{\mathrm{deploy}}$ unless stated; pre-peak rows need a pretrained P3 baseline. Backbones are heterogeneous models, not replicates, and all values are descriptive associations.

man $\rho = 0.79$, versus $r = 0.50$, $\rho = 0.61$ for P3; CUB-trained $+0.72$, NABirds-trained $+0.79$, ImageNet-trained $+0.77$; N $+0.76$, U $+0.65$, SSR $+0.82$), and the P1 and P3 decreases themselves correlate at $r = 0.54$ across cells. Controlling for the peak-to-deployment time gap leaves the P1 association strong within each training domain (partial r 0.74–0.83) and weaker for P3 (0.33–0.62). Held-out P1 is class-name retrieval on held-out species and is not treated as a name-free semantic-capability measure; the association is descriptive and does not imply a shared mechanism for the P1 and P3 changes.

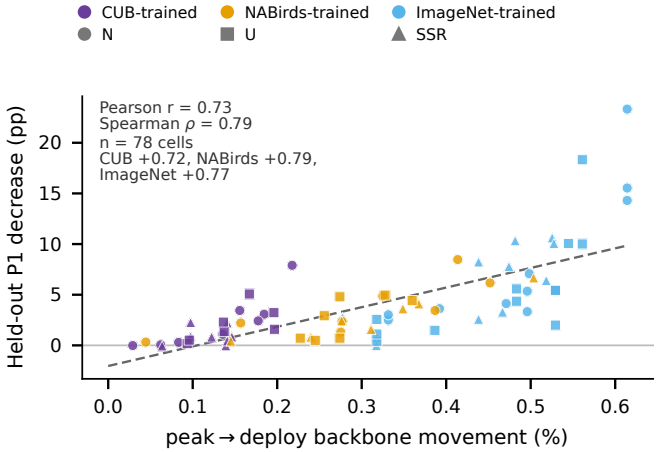


Figure S11: Backbone movement versus held-out P1 decrease, N/U/SSR grid, same 78 held-out cells and same joint image–text backbone movement as the right panel of Figure 5a. P1 is evaluated at each trajectory's P3-peak checkpoint and at its deployment checkpoint; no separate P1-optimal checkpoint is selected, and P1 decrease = P1 (P3-peak checkpoint) − P1 (deployment). This is a robustness analysis; held-out P1 is not treated as a name-free semantic-capability measure.

G.3 Wrong-class dependence of the P3 margin decrease

The Figure 5b analysis follows each image's peak-selected hardest wrong class. The correct anchor is the ground-truth class, and means are class-macro over the 348 classes. Because the competitor is each image's hardest wrong class at the peak and P3 R@1 is low, its mean similarity can exceed that of the ground-truth anchor; the relevant quantity is the paired peak-to-deployment margin change. Mean cosine similarity rises for both the correct attribute-profile anchor and the peak-selected wrong class, but rises more for the competitor. To clarify its scope, we repeated it for wrong classes at predetermined peak similarity ranks (Table S38; hardest = rank 1 of the 347 wrong classes, hard = rank 35, medium = rank 174, easy = rank 312; class identity fixed at the peak and reevaluated at deployment; class-macro, 10,000 class-bootstrap resamples). On P3 the correct-minus-wrong margin decrease is 0.00356 $[0.00194, 0.00518]$ for the hardest, 0.00393 $[0.00227, 0.00553]$ for the hard, 0.00760 $[0.00569, 0.00941]$ for the median, and 0.01238 $[0.01027, 0.01439]$ for the easy wrong class; across 15 rank levels the Spearman correlation between rank and encroachment is $+0.94$ (class-bootstrap 95% [0.89, 0.99]; descriptive, because ranks are nested within images). A split-half selection check (ranking from half of each class's images, evaluating on the other half) gives $+0.0034$, $+0.0047$, $+0.0072$ and $+0.0116$, so the hardest-competitor effect remains positive ($+0.0034$, 95% CI about $[+0.0014, +0.0053]$). The decrease in correct-minus-wrong P3 margin is therefore not concentrated among the wrong classes that are hardest at the peak; it extends across the wrong-class similarity spectrum and is larger for initially easier wrong classes. At the class level the hardest-competitor margin change is heterogeneous (mean -0.00356, 95% CI $[-0.00518, -0.00194]$; median -0.0034; IQR $[-0.0141, +0.0073]$; 10th–90th percentile $[-0.0245, +0.0170]$; 59.5% of classes decrease and 40.5% increase). P1 does not reproduce the monotonic P3 pattern: its encroachment is $+0.0022$ $[+0.0006, +0.0038]$ at rank 1, -0.0033 $[-0.0054, -0.0010]$ at rank 35, $+0.0029$ $[+0.0007, +0.0051]$ at rank 174 and $+0.0083$ $[+0.0060, +0.0107]$ at rank 312 (non-monotone; the negative rank-35 value persists under split-half selection), so P1 and P3 do not share a single competitor-rank dependence.

Probe, level	Rank	Wrong sim. peak→deploy	Δcorr.	Δwrong	Encroachment [95% CI]	% cls. ↓ margin	Split-half [95% CI]
P3 hardest	1	0.203 → 0.211	+0.0044	+0.0080	+0.0036 [+0.0019, +0.0052]	59.5	+0.0034 [+0.0014, +0.0053]
P3 hard	35	0.090 → 0.098	+0.0044	+0.0084	+0.0039 [+0.0023, +0.0055]	62.6	+0.0047 [+0.0024, +0.0070]
P3 medium	174	−0.007 → 0.005	+0.0044	+0.0120	+0.0076 [+0.0057, +0.0094]	67.0	+0.0072 [+0.0049, +0.0095]
P3 easy	312	−0.096 → −0.079	+0.0044	+0.0168	+0.0124 [+0.0103, +0.0144]	74.1	+0.0116 [+0.0088, +0.0141]
P1 hardest	1	0.356 → 0.399	+0.0408	+0.0430	+0.0022 [+0.0006, +0.0038]	56.0	+0.0021 [+0.0004, +0.0039]
P1 hard	35	0.089 → 0.127	+0.0408	+0.0375	−0.0033 [−0.0054, −0.0010]	42.0	−0.0044 [−0.0074, −0.0013]
P1 medium	174	−0.039 → 0.005	+0.0408	+0.0436	+0.0029 [+0.0007, +0.0051]	54.9	+0.0028 [−0.0002, +0.0058]
P1 easy	312	−0.152 → −0.103	+0.0408	+0.0491	+0.0083 [+0.0060, +0.0107]	64.1	+0.0087 [+0.0055, +0.0120]

Table S38: Peak-ranked wrong-class levels for the SSR+ND DFN5B H/14-378 CUB→NAB-348 run. For every image the 347 wrong classes are ranked by similarity at the peak (epoch 20) only (rank 1 = most similar); the class at rank 1, 35 (about the 90th similarity percentile), 174 (median) and 312 (about the 10th percentile) is held fixed and reevaluated at deployment (epoch 60). Encroachment $= \Delta s_{\text{wrong}} - \Delta s_{\text{correct}} = -\Delta\text{margin}$, with margin = correct − wrong cosine similarity; class-macro means over the 348 NAB-348 classes, paired class bootstrap (10,000 resamples). The split-half column ranks wrong classes from half of each class's images and evaluates on the other half. P1 and P3 are separate capabilities and are not pooled.

G.4 Encoder-state swap

Table S39 gives the full swap results for the representative run. Held-out NAB-348 P3 is 17.63 at peak and 14.50 at deployment; combining the deployment image encoder with the peak text/routing encoder gives 15.15, and the reverse gives 17.09. Table S40 gives the paired class-bootstrap de-

State	Image encoder	Text/routing encoder	CUB P1	CUB P3	NAB-348 P1	NAB-348 P3
peak	ep. 20	ep. 20	90.02	45.56	76.10	17.63
image late only	ep. 60	ep. 20	89.97	45.20	71.23	15.15
text late only	ep. 20	ep. 60	89.97	47.07	76.07	17.09
deploy	ep. 60	ep. 60	90.23	46.29	70.81	14.50

Table S39: Encoder-state swap, DFN5B H/14-378 SSR+ND fine-tuned on CUB-200 (single run). Image encoder and text/routing encoder are taken from epoch 20 (peak) or epoch 60 (deployment); "image late only" combines the epoch-60 image encoder with the epoch-20 text/routing encoder, "text late only" the reverse. Values are R@1 (%).

composition (NAB-348 P3: image-late -2.53 pp [-3.53, -1.56], text/SSR-late -0.66 pp [-1.19, -0.15], difference -1.88 pp [-2.84, -0.90], interaction approximately zero). This localization is specific to this representative transition.

G.5 Additional encoder-state swaps on other SSR+ND runs

To test whether the representative localization above recurs, we ran the same evaluation-only four-state encoder swap (same encoder-state substitution procedure and encoder partition: image backbone; text backbone plus SSR modules, with the logit scale and buffers taken from the deployment base checkpoint) on three additional SSR+ND runs, with no new training. The selection rule was fixed before the encoder-swap analysis. Eligible runs were identified from the NAB-trained SSR+ND runs audited on the dataset-class-disjoint CUB-59 subset (all 18 saved checkpoints; own CUB-59 P3 peak, epochs 4/8/2/4 for DFN5B H/14-378, DFN2B L/14, MetaCLIP H/14, MetaCLIP L/14; all four had a peak before the P1-selected deployment checkpoint and a decrease above 0.5 pp), and candidates were then taken in a fixed order: MetaCLIP H/14 CUB→NAB-348 (peak 12, deployment 75; only a 0.64 pp decline, retained because it is the only different-backbone SSR+ND case of the same strict transfer), DFN2B L/14 NABirds→CUB-59 (8, 45) and MetaCLIP L/14 NABirds→CUB-59 (4, 40). DFN5B NABirds→CUB-59 (which had the largest decline, 13.76 pp) was not selected because it was not among the candidates under this rule; runs were selected by the fixed rule, not by decline magnitude or any swap result. Results are in Table S41.

The outcome is heterogeneous and largely unresolved. For MetaCLIP H/14 (CUB→NAB-348), the image-late effect on held-out P3 is negative (-0.73 pp, CI including zero) and the text/SSR-late

Eval.	Probe	Effect	Change (pp) [95% CI]
CUB-200	P3	image-late	−0.44 [−1.84, +0.89]
CUB-200	P3	text/SSR-late	+1.45 [+0.55, +2.37]
CUB-200	P3	total	+0.59 [−1.06, +2.18]
CUB-200	P3	interaction	−0.42 [−1.15, +0.29]
CUB-200	P3	image-late − text/SSR-late	−1.90 [−3.36, −0.44]
NAB-348	P3	image-late	−2.53 [−3.53, −1.56]
NAB-348	P3	text/SSR-late	−0.66 [−1.19, −0.15]
NAB-348	P3	total	−3.21 [−4.49, −2.03]
NAB-348	P3	interaction	−0.02 [−0.52, +0.46]
NAB-348	P3	image-late − text/SSR-late	−1.88 [−2.84, −0.90]
NAB-348	P1	image-late	−4.94 [−6.13, −3.79]
NAB-348	P1	text/SSR-late	−0.04 [−0.51, +0.45]
NAB-348	P1	total	−5.28 [−6.56, −4.01]
NAB-348	P1	interaction	−0.30 [−0.77, +0.15]
NAB-348	P1	image-late − text/SSR-late	−4.90 [−6.12, −3.74]
CUB-200	P1	image-late	−0.08 [−0.65, +0.49]
CUB-200	P1	text/SSR-late	−0.04 [−0.70, +0.64]
CUB-200	P1	total	+0.23 [−0.45, +0.95]
CUB-200	P1	interaction	+0.36 [−0.14, +0.85]
CUB-200	P1	image-late − text/SSR-late	−0.03 [−0.78, +0.73]

Table S40: Paired class-bootstrap decomposition of the encoder-state swap (Table S39); representative DFN5B H/14-378 SSR+ND CUB-trained run. "Image-late" is the effect of substituting the deployment image encoder at fixed peak text/routing state, "text/SSR-late" the reverse, "interaction" the residual of the additive decomposition. Class bootstrap over evaluation classes (10,000 resamples).

effect is near zero, the same direction as the representative run but not resolved; on held-out P1 the image-late effect (-2.44 pp) excludes zero and the image−text contrast excludes zero. For both NABirds→CUB-59 runs, held-out P3 losses are split about equally between the encoders (image-late -0.56 and -0.69 pp, text/SSR-late -0.63 and -0.73 pp; image−text $+0.07$ and $+0.04$ pp) with intervals that span zero over only 59 evaluation classes, so no encoder-specific localization is supported there. Target-domain P1 and P3 are not stable across the cross states in these runs, because their P3 peaks are early relative to deployment and target performance is still changing. We therefore keep the encoder localization explicitly representative of the single DFN5B/CUB→NAB-348 run and do not generalize it. These are independent existing runs, not stochastic replicates, and the intervals reflect evaluation-class uncertainty only.

G.6 Cross-state P1/P3 margin recombination

Table S42 gives raw correct similarity, hardest-competitor similarity, and margin for the four states; these results are plotted in Figure S12b. They agree with the encoder swap: substituting only the late image state reproduces nearly all of the held-out margin loss for P3 and P1.

G.7 Correct-anchor versus hardest-competitor similarity

Table S43 (plotted in Figure S12a) reports the per-checkpoint hardest-wrong version: from peak to deployment the correct-class similarity is roughly flat to slightly rising while the hardest competitor's similarity rises more, for both SSR+ND and the text-frozen U run.

G.8 Fixed-competitor encroachment

Table S44 gives the peak→deployment competitor-encroachment analysis summarized in §4.4 and plotted per class in Figure 5b, with per-checkpoint (dynamic) and fixed-at-peak competitors, on CUB-200 and NAB-348 and for P1 and P3. For the SSR+ND run on NAB-348 P3, the encroachment decomposes into a peak-to-deployment cosine-similarity rise of $+0.00443$ (95% class-bootstrap CI

	Class-macro R@1 (%)				Effect (pp) [95% CI]			
Evaluation	P_I+P_T	D_I+P_T	P_I+D_T	D_I+D_T	Image-late	Text/SSR-late	Image−text	Interaction
MetaCLIP H/14, CUB→NAB-348 (ep. 12/75)								
held-out, NAB-348 P3	12.39	11.66	12.46	11.64	−0.73 [−1.54, +0.08]	+0.06 [−0.38, +0.50]	−0.80 [−1.68, +0.10]	−0.08 [−0.52, +0.34]
held-out, NAB-348 P1	72.12	69.67	72.20	70.37	−2.44 [−3.62, −1.30]	+0.08 [−0.58, +0.74]	−2.52 [−3.82, −1.25]	+0.61 [+0.12, +1.11]
target, CUB-200 P3	38.35	40.35	39.33	42.01	+2.00 [+0.51, +3.53]	+0.98 [−0.01, +1.95]	+1.02 [−0.32, +2.35]	+0.68 [−0.26, +1.69]
target, CUB-200 P1	83.30	85.87	85.16	87.95	+2.56 [+1.40, +3.77]	+1.85 [+0.66, +3.13]	+0.71 [−0.90, +2.33]	+0.23 [−0.65, +1.14]
DFN2B L/14, NABirds→CUB-59 (ep. 8/45)								
held-out, CUB-59 P3	46.35	45.79	45.72	44.42	−0.56 [−2.76, +1.59]	−0.63 [−2.38, +0.97]	+0.07 [−2.52, +2.69]	−0.74 [−2.27, +0.84]
held-out, CUB-59 P1	87.69	87.98	88.40	87.96	+0.29 [−0.77, +1.40]	+0.71 [−0.53, +2.03]	−0.42 [−1.38, +0.52]	−0.74 [−1.52, +0.02]
target, NAB P3	8.11	8.63	8.00	8.49	+0.52 [−0.04, +1.08]	−0.11 [−0.45, +0.25]	+0.63 [−0.04, +1.31]	−0.03 [−0.28, +0.23]
target, NAB P1	78.78	79.66	80.58	81.70	+0.88 [+0.10, +1.65]	+1.80 [+0.90, +2.72]	−0.92 [−2.03, +0.15]	+0.25 [−0.34, +0.81]
MetaCLIP L/14, NABirds→CUB-59 (ep. 4/40)								
held-out, CUB-59 P3	47.43	46.74	46.70	45.49	−0.69 [−4.30, +3.14]	−0.73 [−3.77, +2.47]	+0.04 [−4.48, +4.62]	−0.52 [−2.67, +1.54]
held-out, CUB-59 P1	83.62	82.17	82.98	80.51	−1.45 [−4.27, +1.34]	−0.64 [−3.30, +1.78]	−0.80 [−3.92, +2.60]	−1.02 [−2.79, +0.73]
target, NAB P3	9.41	10.76	10.78	12.67	+1.35 [+0.47, +2.23]	+1.37 [+0.71, +2.06]	−0.02 [−1.08, +1.06]	+0.54 [−0.02, +1.11]
target, NAB P1	70.79	76.15	76.36	81.97	+5.36 [+4.20, +6.54]	+5.58 [+4.18, +7.03]	−0.21 [−1.82, +1.41]	+0.24 [−0.66, +1.15]

Table S41: Additional encoder-state swaps on three additional SSR+ND runs, selected by a rule fixed before any swap was run (Appendix G.5). States: P = P3-peak checkpoint, D = deployment checkpoint, I = image encoder, T = text/SSR state (checkpoint epochs peak/deploy in the run label); effects use the same definitions as Table S40 (image-late = D_I+P_T − P_I+P_T; text/SSR-late = P_I+D_T − P_I+P_T). Paired class bootstrap over evaluation classes (10,000 resamples); NABirds-trained runs are evaluated on the dataset-class-disjoint CUB-59 subset (59 classes), target rows on the run's own training-domain test set. Independent existing runs, not stochastic replicates.

State	P1 correct	P1 wrong	P1 margin	P3 correct	P3 wrong	P3 margin
CUB-200 (n=200)						
peak	0.4533	0.3183	0.1349	0.2157	0.2280	-0.0123
image late only	0.5089	0.3453	0.1636	0.2352	0.2485	-0.0134
text late only	0.4567	0.3064	0.1504	0.2117	0.2201	-0.0084
deploy	0.5232	0.3366	0.1866	0.2327	0.2416	-0.0089
NAB-348 (n=348)						
peak	0.4077	0.3559	0.0518	0.1383	0.2028	-0.0645
image late only	0.4492	0.4042	0.0450	0.1467	0.2228	-0.0761
text late only	0.4069	0.3554	0.0515	0.1335	0.1985	-0.0650
deploy	0.4484	0.4038	0.0446	0.1427	0.2197	-0.0770

Table S42: Cross-state embedding recombination, SSR+ND (DFN5B H/14-378, CUB-trained; single run). For each class, "correct" is the image-to-text cosine similarity to the correct class text anchor, "wrong" is the similarity to the hardest competing class anchor, and margin is correct minus wrong; entries are class-macro averages over the stated number of classes, computed for P1 and P3 text anchors. States follow Table S39. Figure 5c plots the NAB-348 margins as change relative to the peak state.

Model	Eval.	Anchor	PT corr.	PT wrong	Peak corr.	Peak wrong	Dep. corr.	Dep. wrong
SSR+ND	CUB-200	P3	0.3100	0.3217	0.2157	0.2280	0.2327	0.2416
	CUB-200	P1	0.3659	0.3290	0.4533	0.3183	0.5232	0.3366
	NAB-348	P3	0.2796	0.3082	0.1383	0.2028	0.1427	0.2197
	NAB-348	P1	0.3669	0.3425	0.4077	0.3559	0.4484	0.4038
U (text-frozen)	CUB-200	P3	0.3100	0.3217	0.1993	0.2109	0.2648	0.2462
	CUB-200	P1	0.3659	0.3290	0.4017	0.3157	0.4669	0.3386
	NAB-348	P3	0.2796	0.3082	0.1325	0.1792	0.1519	0.2109
	NAB-348	P1	0.3669	0.3425	0.3872	0.3457	0.4129	0.3835

Table S43: Correct-class versus hardest-competitor text-anchor similarity (class-macro means) at pretrained (PT), peak, and deployment checkpoints for SSR+ND (peak epoch 20, deployment epoch 60) and an independent text-frozen U run (peak epoch 20, deployment epoch 75; a separate run from the main U run, whose deployment epoch is 25). In the text-frozen run the entire text encoder including layer norms is locked, and the text anchors were verified to be bit-identical across PT, peak, and deployment, so all change is image-side.

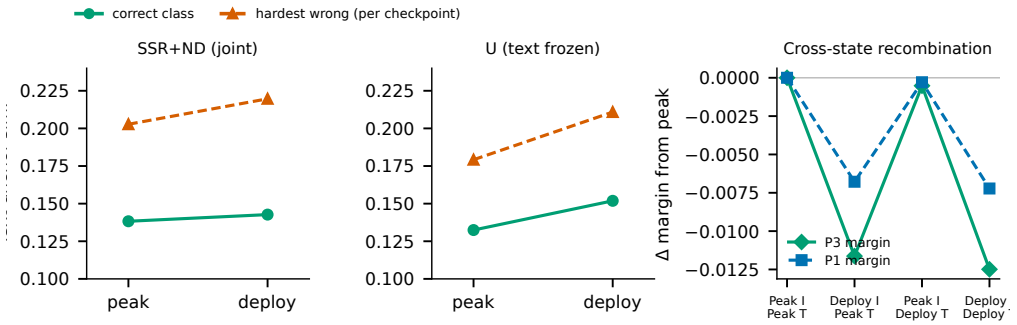


Figure S12: Supplementary late-specialization diagnostics. **(a)** Correct-class versus hardest-wrong text-anchor similarity, peak to deployment, where the hardest wrong class is re-selected at each checkpoint (see Table S44 for why this statistic can overstate encroachment). **(b)** Change in NAB-348 correct-minus-competitor margin from peak under cross-state recombination (I = image encoder, T = text/routing encoder).

$[+0.00227, +0.00665]$) for the correct attribute-profile class and $+0.00799$ ($[+0.00629, +0.00973]$) for the same peak-selected wrong class, a paired difference of $+0.00356$ ($[+0.00194, +0.00518]$); both similarities rise, the competitor's more, so the correct-minus-wrong margin decreases by 0.00356 (Figure 5b). On NAB-348 P3 the fixed-competitor encroachment is resolved and positive for SSR+ND ($+3.6 \times 10^{-3}$, 59% of classes positive) and consistent with zero for the independent text-frozen U control (-1.3×10^{-3}, 47%); the dynamic hardest-wrong statistic is positive for both, illustrating why it is not used as the primary metric. On NAB-348 P1 the fixed-competitor effect is positive for both models, and on the CUB target it is negative for both (the correct anchor gains most); P1 and P3 are distinct capabilities and are not pooled. Table S45 reports the class-level relationship between encroachment and the change in class top-1 accuracy, which is negative in every tested model×dataset×probe cell but partly mechanical. A separate semantic-neighborhood analysis of the same saved embeddings (pretrained text-text similarity freezes neighbor membership; class bootstrap) found that acquisition and late specialization have opposite-signed relationships to pretrained semantic distance, and that held-out errors remain concentrated among pretrained semantic neighbors (rank-1 neighbor: 36.2% of late-phase flips for U with text frozen on NAB-348 P1 against a 0.29% chance baseline); P1 and P3 do not show identical held-out late-phase neighbor dynamics, and no single neighbor mechanism is claimed.

G.9 Text-frozen control

In the text-frozen U run the entire text encoder (including layer norms) is locked while the image encoder and logit scale are trained. The text anchors were verified to be bit-identical across PT, peak, and deployment, so every change in Table S43 for this model is image-side by construction. This run has peak epoch 20 and deployment epoch 75; it is a separate run from the main U run (deployment epoch 25).

G.10 Image-only oracle prototype diagnostic

Table S46 reports leave-one-out image-prototype accuracy (an oracle using held-out-image class prototypes, not a zero-shot capability): about 88–91% across PT, peak, and deployment. Class structure is largely preserved, though NAB-348 declines by 2.7–3.1 pp from peak to deployment.

G.11 Geometry hypotheses and null results

Table S47 reports all coarse-geometry correlations, weak/non-null and null alike. Tangent-residual alignment relates weakly but measurably to ΔP3 (all four intervals exclude 0, $\rho \in [0.15, 0.33]$). Residual-direction coherence, manifold radius, and effective rank show no reliable relationship with

Model	Eval.	Probe	Competitor	Encroachment ($\times 10^{-3}$) [95% CI]	Classes > 0
SSR+ND (joint)	CUB-200	P1	hardest wrong per checkpoint	-51.66 [-55.49, -47.88]	2%
SSR+ND (joint)	CUB-200	P1	fixed at peak (per image)	-63.80 [-67.61, -60.04]	1%
SSR+ND (joint)	CUB-200	P1	fixed at peak (per class)	-71.66 [-75.93, -67.35]	2%
SSR+ND (joint)	CUB-200	P3	hardest wrong per checkpoint	-3.41 [-6.03, -0.85]	40%
SSR+ND (joint)	CUB-200	P3	fixed at peak (per image)	-11.84 [-14.41, -9.35]	22%
SSR+ND (joint)	CUB-200	P3	fixed at peak (per class)	-11.34 [-14.31, -8.40]	28%
SSR+ND (joint)	NAB-348	P1	hardest wrong per checkpoint	$+7.21$ [+5.55, +8.88]	70%
SSR+ND (joint)	NAB-348	P1	fixed at peak (per image)	$+2.22$ [+0.60, +3.82]	56%
SSR+ND (joint)	NAB-348	P1	fixed at peak (per class)	$+2.12$ [+0.31, +3.88]	57%
SSR+ND (joint)	NAB-348	P3	hardest wrong per checkpoint	$+12.49$ [+10.80, +14.12]	76%
SSR+ND (joint)	NAB-348	P3	fixed at peak (per image)	$+3.56$ [+1.94, +5.18]	59%
SSR+ND (joint)	NAB-348	P3	fixed at peak (per class)	$+3.42$ [+1.43, +5.42]	55%
U (text frozen)	CUB-200	P1	hardest wrong per checkpoint	-42.33 [-45.97, -38.68]	6%
U (text frozen)	CUB-200	P1	fixed at peak (per image)	-50.74 [-54.61, -46.86]	3%
U (text frozen)	CUB-200	P1	fixed at peak (per class)	-51.53 [-55.22, -47.82]	2%
U (text frozen)	CUB-200	P3	hardest wrong per checkpoint	-30.21 [-33.84, -26.57]	12%
U (text frozen)	CUB-200	P3	fixed at peak (per image)	-42.14 [-45.98, -38.34]	6%
U (text frozen)	CUB-200	P3	fixed at peak (per class)	-41.83 [-46.10, -37.63]	8%
U (text frozen)	NAB-348	P1	hardest wrong per checkpoint	$+12.06$ [+10.19, +14.01]	76%
U (text frozen)	NAB-348	P1	fixed at peak (per image)	$+4.88$ [+3.09, +6.68]	61%
U (text frozen)	NAB-348	P1	fixed at peak (per class)	$+5.83$ [+3.89, +7.81]	64%
U (text frozen)	NAB-348	P3	hardest wrong per checkpoint	$+12.26$ [+10.38, +14.20]	74%
U (text frozen)	NAB-348	P3	fixed at peak (per image)	-1.31 [-3.21, +0.67]	47%
U (text frozen)	NAB-348	P3	fixed at peak (per class)	-1.09 [-3.42, +1.33]	45%

Table S44: Peak→deployment competitor encroachment, $\Delta s_{\text{wrong}} - \Delta s_{\text{correct}}$ (class-macro mean of image–text-anchor cosine changes; positive = the competitor closes the gap relative to the correct anchor). "Hardest wrong per checkpoint" re-selects the maximizing wrong class independently at peak and at deployment and can rise merely because wrong-class similarities spread or the maximizer changes; the fixed-competitor rows freeze each image's (or class's) hardest wrong class at the peak checkpoint and re-evaluate it at deployment. Intervals: paired class bootstrap over evaluation classes (10,000 resamples), not training-seed replication. Peak epoch 20; deployment epoch 60 (SSR+ND) or 75 (U, text encoder frozen; text anchors bit-identical across checkpoints). Single runs, DFN5B H/14-378 fine-tuned on CUB-200.

Model	Eval.	Probe	Competitor	Spearman ρ [95% CI]	Mean Δacc. (pp)
SSR+ND (joint)	CUB-200	P1	hardest wrong per checkpoint	-0.09 [-0.24, +0.06]	+0.2
SSR+ND (joint)	CUB-200	P1	fixed at peak (per image)	-0.09 [-0.24, +0.06]	+0.2
SSR+ND (joint)	CUB-200	P3	hardest wrong per checkpoint	-0.60 [-0.69, -0.51]	+0.6
SSR+ND (joint)	CUB-200	P3	fixed at peak (per image)	-0.57 [-0.65, -0.47]	+0.6
SSR+ND (joint)	NAB-348	P1	hardest wrong per checkpoint	-0.57 [-0.63, -0.49]	-5.3
SSR+ND (joint)	NAB-348	P1	fixed at peak (per image)	-0.58 [-0.64, -0.50]	-5.3
SSR+ND (joint)	NAB-348	P3	hardest wrong per checkpoint	-0.39 [-0.48, -0.30]	-3.2
SSR+ND (joint)	NAB-348	P3	fixed at peak (per image)	-0.35 [-0.44, -0.26]	-3.2
U (text frozen)	CUB-200	P1	hardest wrong per checkpoint	-0.20 [-0.34, -0.05]	+1.4
U (text frozen)	CUB-200	P1	fixed at peak (per image)	-0.17 [-0.32, -0.02]	+1.4
U (text frozen)	CUB-200	P3	hardest wrong per checkpoint	-0.36 [-0.48, -0.22]	+16.6
U (text frozen)	CUB-200	P3	fixed at peak (per image)	-0.39 [-0.51, -0.25]	+16.6
U (text frozen)	NAB-348	P1	hardest wrong per checkpoint	-0.64 [-0.70, -0.56]	-9.9
U (text frozen)	NAB-348	P1	fixed at peak (per image)	-0.60 [-0.66, -0.52]	-9.9
U (text frozen)	NAB-348	P3	hardest wrong per checkpoint	-0.45 [-0.53, -0.36]	-2.6
U (text frozen)	NAB-348	P3	fixed at peak (per image)	-0.41 [-0.49, -0.32]	-2.6

Table S45: Class-level Spearman correlation between encroachment and the peak→deployment change in class top-1 accuracy (class bootstrap over evaluation classes). The relationship is partly mechanical, because encroachment and top-1 retrieval derive from the same similarity scores; it is supportive consistency, not independent mechanistic evidence.

Model	Eval.	PT	Peak	Deploy
SSR+ND	CUB-200	89.2	90.4	90.6
	NAB-348	90.9	90.6	88.4
U (text-frozen)	CUB-200	89.2	90.3	90.6
	NAB-348	90.8	91.2	88.0

Table S46: Image-only leave-one-out prototype accuracy (class-macro, %). For each image, the class prototype is the mean image embedding of the other images of the same class; the image is assigned to the nearest prototype. This is an oracle diagnostic that uses held-out-image prototypes; it is not a zero-shot capability. Class structure remains largely preserved (88.0–91.2%), although NAB-348 shows a 2.7–3.1 pp decline from peak to deployment.

ΔP3; some show non-null relationships with ΔP1, so the null statement applies to ΔP3 only. No coarse geometry metric provides a complete explanation.

Geometry metric	Target	SSR+ND CUB	SSR+ND NAB	U-tf CUB	U-tf NAB
Tangent-residual alignment (P3)	ΔP3	0.33*	0.25*	0.15*	0.21*
	ΔP1	0.05	-0.09	-0.04	0.00
Tangent-residual alignment (P2)	ΔP3	0.05	0.11*	0.14	0.19*
	ΔP1	0.01	0.05	0.03	0.10
Raw alignment (P3)	ΔP3	0.23*	0.19*	0.12	0.19*
	ΔP1	0.02	-0.15*	-0.06	-0.08
Residual-direction coherence	ΔP3	-0.04	-0.06	-0.05	-0.07
	ΔP1	0.02	0.21*	0.04	0.13*
Manifold radius	ΔP3	-0.03	0.06	0.06	0.06
	ΔP1	-0.05	-0.26*	-0.07	-0.20*
Effective rank	ΔP3	0.08	-0.01	0.04	-0.02
	ΔP1	0.10	0.11*	0.17*	0.02
Participation ratio	ΔP3	0.07	0.00	0.03	-0.02
	ΔP1	–	–	–	–
PC1 variance fraction	ΔP3	-0.04	0.01	-0.00	0.02
	ΔP1	–	–	–	–

Table S47: Class-level Spearman correlations between late-minus-peak changes in coarse geometry metrics and changes in class accuracy (n = 200 CUB classes, 348 NAB classes; 10,000 class resamples for the 95% intervals). Asterisks mark intervals excluding 0. Tangent-residual alignment vs. ΔP3 is weak but non-null in all four cells ($\rho \in [0.15, 0.33]$); residual-direction coherence, manifold radius, and effective rank show no reliable relationship with ΔP3, although some show non-null relationships with ΔP1. The two models are SSR+ND and an independent text-frozen U run (U-tf). Full intervals are provided in the supporting results.

H Additional Controls and Alternative Interventions

H.1 WiSE interpolation

Weights are interpolated as $\theta(\alpha) = (1-\alpha)\theta_{\mathrm{PT}} + \alpha\theta_{\mathrm{FT}}$ at evaluation time for N (deployment epoch 40) and U (deployment epoch 25), DFN5B H/14-378 fine-tuned on CUB-200 (Table S48). At $\alpha = 0.5$, held-out NAB-348 P3 recovers toward pretrained while target CUB P1/P3 remain close to or above the deployed values (N: CUB P1 91.5 vs. 91.1 deployed, CUB P3 36.0 vs. 31.1, NAB P3 15.1 vs. 13.7, NAB P1 82.3 vs. 71.6; U: CUB P1 91.5 vs. 91.2, CUB P3 41.7 vs. 35.8, NAB P3 17.0 vs. 15.1, NAB P1 82.0 vs. 76.0). Moving weights toward the pretrained model can therefore recover substantial held-out capability while preserving strong target performance in this representative setting, so specialization need not be irreversible. For context, WiSE-interpolated U

at $\alpha = 0.5$ is comparable to or better than the deployed SSR run on these two P3 metrics (SSR: CUB 40.4, NAB 16.9), whereas SSR+ND has higher target P3 (46.3) and lower NAB P3 (14.5). WiSE is a probe of reversibility, not a competing method. SSR and SSR+ND are not evaluated because their special tokens and gates have no direct pretrained counterparts. Single seed, coarse grid, one backbone/domain.

Method	α	CUB P1	CUB P3	NAB-348 P1	NAB-348 P3
N	0	88.83	34.35	82.93	13.85
	0.25	90.87	36.61	83.60	14.60
	0.5	91.46	35.95	82.25	15.06
	0.75	91.39	33.62	78.06	15.00
	1	91.06	31.12	71.55	13.74
U	0	88.75	34.45	83.03	13.87
	0.25	90.82	39.01	83.25	15.57
	0.5	91.47	41.68	82.03	17.00
	0.75	91.49	40.30	79.97	16.95
	1	91.16	35.78	76.00	15.14

Table S48: WiSE-style weight interpolation $\theta(\alpha) = (1-\alpha)\theta_{\mathrm{PT}} + \alpha\theta_{\mathrm{FT}}$ for N (deployment epoch 40) and U (deployment epoch 25), DFN5B H/14-378 fine-tuned on CUB-200; $\alpha = 0$ is the pretrained model and $\alpha = 1$ the deployed checkpoint. Interpolation is applied at evaluation time to all trainable parameters; no training is involved. Values are R@1 (%) on CUB-200 (target) and NAB-348 (strict held-out). Single seed, coarse α grid, one backbone and training domain.

H.2 Hard-negative objective

Motivation. SSR improves name-free semantic capability without explicitly training the model to separate a correct attribute profile from nearby semantic alternatives. This control asks whether directly strengthening such target-side discrimination is enough to improve transferable semantic accessibility, or whether it can prevent the later loss of transferable P3. Starting from the canonical SSR recipe, we add a controlled hard-negative ranking term in which the negative text differs from the correct profile by a single attribute value. This is a diagnostic intervention, not a proposed additional method.

Construction and objective. For a CUB training image with its correct structured profile (name, concept and original attributes; the *correct* or *positive* profile), the controlled semantic hard negative keeps the same name, concept and remaining attributes and changes exactly one attribute phrase by a minimal edit (a single colour, shape or size change). The negative pool contains 618 valid candidates covering all 200 CUB classes. During training one negative is sampled per class at each effective optimizer step, shared by all images of that class in the step and resampled at the next step. Let s_{pos} and s_{neg} be the (logit-scaled) image–text similarities of the image to the final fused text representation of its positive profile and of its hard negative. The auxiliary loss is

$$L_{\mathrm{HN}} = \mathrm{mean}\,\mathrm{softplus}(s_{\mathrm{neg}} - s_{\mathrm{pos}}), \qquad L = L_{\mathrm{CLIP}} + \lambda L_{\mathrm{HN}}, \quad \lambda = 0.5, \tag{2}$$

where L_{CLIP} is the standard multi-positive contrastive loss of the SSR recipe. The penalty is small when $s_{\mathrm{pos}} > s_{\mathrm{neg}}$ and grows as the negative approaches or exceeds the positive, so the term directly encourages positive-versus-negative separation. The run is DFN5B H/14-378, fine-tuned on CUB-200 only, one seed, one λ; the matched baseline is the canonical SSR recipe run from the identical starting checkpoint with the same epoch schedule. The target is CUB-200 and the strict held-out population is NAB-348.

Matched-epoch result. Target CUB P3 is improved at later checkpoints and reaches a higher peak (43.04 at epoch 40 versus the SSR peak of 40.42 at epoch 20), and target image–P3 alignment is higher (0.355 versus 0.315 at epoch 100). Strict held-out NAB-348 P3 does not show the same improvement at matched evaluated epochs: from epoch 8 onward it is below the SSR baseline, and by epoch 100 it is 12.38 versus 13.30 despite the stronger target-side discrimination (held-out

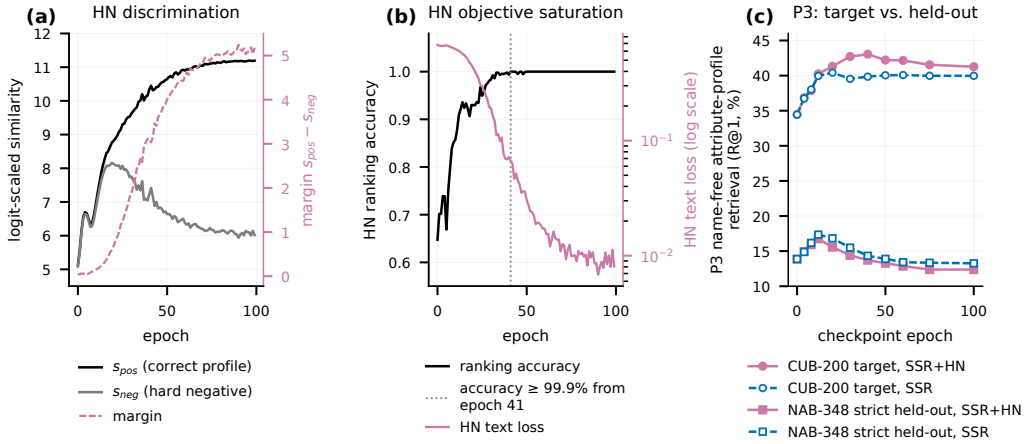


Figure S13: **Hard-negative semantic-discrimination control,** DFN5B H/14-378 fine-tuned on CUB-200 (single seed, $\lambda = 0.5$). Starting from SSR, a target-side ranking loss separates the correct attribute profile from a controlled hard negative differing in one attribute value. **(a)** Logit-scaled similarities of the image to the positive (s_{pos}) and hard-negative (s_{neg}) profiles and their margin, per training epoch (training-log quantities on the hard-negative batches, not an evaluation probe): s_{neg} first rises (until about epoch 19) and then declines while s_{pos} keeps increasing, so the relative margin grows. **(b)** Hard-negative ranking accuracy (fraction with $s_{\text{pos}} > s_{\text{neg}}$) first reaches 99.9% at epoch 41 (dotted line) and stays there, while the hard-negative text loss keeps decreasing. **(c)** P3 (name-free attribute-profile retrieval) on CUB-200 (target) and strict held-out NAB-348 at the 11 evaluated checkpoints (no interpolation): the hard-negative run exceeds the matched SSR baseline on target CUB P3 at later checkpoints but not on held-out NAB-348 P3 at matched epochs. This representative control shows that stronger target-side semantic discrimination need not translate into stronger transferable P3.

alignment is very similar, 0.231 versus 0.228). Held-out P2 is consistently lower for the hard-negative run, and P1 is slightly lower at matched epochs (-0.1 to -1.6 pp). Exact values at five of the evaluated epochs are in Table S49. (The table's SSR baseline comes from the alignment evaluation and differs from the trajectory evaluation plotted in Figure S13c by at most about 0.05 pp, e.g. NAB-348 P3 at epoch 100 is 13.30 versus 13.26; signs are unchanged.)

Checkpoint caveat. The conclusion depends on the comparison rule. At matched training epochs, stronger target discrimination does not produce a corresponding held-out P3 gain. At each run's own P1-selected deployment checkpoint, however, the hard-negative run is higher on NAB-348 P3 (15.57 versus 13.34) because its deployment occurs much earlier (epoch 20 versus epoch 75 for the matched SSR baseline in this evaluation series). We therefore use this experiment only to show that stronger target-side discrimination is not by itself sufficient to guarantee improved transfer at a matched stage of training.

Interpretation and scope. The control reinforces the capability-trajectory view: optimization can continue to strengthen semantic discrimination on the training ontology while held-out name-free retrieval has already passed its optimum (target CUB P3 stays strong or improves later, whereas NAB-348 P3 peaks early and declines). It is separate from the fixed-competitor analysis in §4.4 and is not a mechanistic explanation of it: the hard-negative objective acts on target CUB semantic alternatives, whereas Figure 5b analyzes the geometry of strict held-out NAB-348 P3 after its peak. The evidence is one backbone, one training domain, one seed and one λ in a diagnostic control, not a method sweep; it does not show that hard negatives fail, hurt transfer, or cause late specialization, and the safe conclusion is only that stronger target-side discrimination need not improve transferable P3 at a matched training stage.

Epoch	CUB P3	NAB-348 P3	CUB P1	CUB align.	NAB align.
8	37.87 / 38.02	15.91 / 16.18	88.28 / 88.44	0.463 / 0.460	0.263 / 0.262
12	40.25 / 39.87	16.76 / 17.32	88.99 / 89.02	0.469 / 0.457	0.287 / 0.283
20	41.32 / 40.42	15.57 / 16.90	90.06 / 90.30	0.454 / 0.412	0.281 / 0.268
40	43.04 / 39.82	13.71 / 14.36	89.75 / 89.94	0.410 / 0.352	0.254 / 0.244
100	41.27 / 40.01	12.38 / 13.30	90.02 / 90.25	0.355 / 0.315	0.231 / 0.228

Table S49: Hard-negative control versus the matched SSR baseline, DFN5B H/14-378, CUB-trained (single seed, single λ). Each cell is hard-negative / SSR. The objective adds λL_{HN} with $\lambda = 0.5$ to the multi-positive contrastive loss, where $L_{\mathrm{HN}} = \mathrm{mean}\,\mathrm{softplus}(s_{\mathrm{neg}} - s_{\mathrm{pos}})$ is computed on the final fused text representation, and each negative differs from the true profile by one controlled attribute edit (pool of 618 candidates covering all 200 classes). Alignment is representational similarity between image and P3 text geometry. At each run's own P1-selected checkpoint (hard-negative epoch 20, SSR epoch 75) CUB P3 is 41.32 vs. 39.97 and NAB-348 P3 is 15.57 vs. 13.34; at epoch 100, NAB-348 P3 is 12.38 vs. 13.30. Peak NAB-348 P3 is 16.76 (hard-negative) vs. 17.35 (SSR).

H.3 Direct name-access ablation: interpretation

See §A.5 and §E.7. The name-access effect of the direct name-access ablation and the late specialization analyzed in §G are mechanistically distinct and should not be conflated.

H.4 Other verified controls

No further controls beyond those in this appendix are reported.

I Qualitative Probe Examples

The examples in Table S50 illustrate the probes; they are not performance claims and were not selected by model success or failure. Dataset photographs are not reproduced because the underlying images are third-party works. Reviewed CAPP profiles are available for CUB and NABirds, so no CAPP example is given for FineView or ImageNet.

Item	CUB: Kentucky Warbler	NAB-348: adult male Lesser Goldfinch
P1	a photo of a Kentucky Warbler	a photo of a adult male Lesser Goldfinch
P3 template	a bird with {attributes}	a bird with {attributes}
CAPP correct profile	a bird with black crown, bright yellow underparts, olive-green upperparts	a bird with bright yellow belly, black cap, glossy cap, pointed wing, white wing patch
CAPP one-slot counterfactual	a bird with black crown, **gray underparts**, olive-green upperparts (slot: underparts color)	a bird with bright yellow belly, black cap, glossy cap, **rounded wing**, white wing patch (slot: wing shape)
Item	FineView: Monarch	ImageNet: Goldfish
P3 template	a butterfly with {attributes}	a object with {attributes}
Attributes	orange upperside; wide black borders; black veins; blurred black veins; white border spots; white wing-tip spots	orange-red body; bright yellow body; torpedo-shaped body; large reflective scales; bifurcated caudal fin; stiff dorsal fin; smooth operculum
CAPP	not available	not available

Table S50: Verified probe examples. The CAPP pairs are drawn from the reviewed full-class CAPP profiles (donor values for the swapped slot come from other classes). Only the changed slot is highlighted. The Figure 1c CAPP panel is an illustrative profile, not one of these verified pairs.

J Reproducibility and Result Provenance

J.1 Full statement of limitations

The main-paper Limitations paragraph is a summary of the following full statement.

Scope and limitations. The fixed-competitor and image-only analyses in §4.4 are single-run diagnostics, and the broader movement result is a descriptive association over four backbones and N/U/SSR (SSR+ND not included), not a causal mechanism. Movement also tracks held-out P1 decrease more strongly over the same P3-peak→deployment intervals (Appendix G.2), so it should be interpreted as accompanying broader held-out specialization rather than as a P3-specific mechanism; heterogeneous encoder-state swaps (Appendix G.5) likewise support no encoder-specific attribution. The six backbones are diverse rather than statistically representative or stochastic replicates, so six-backbone summaries are descriptive and carry no inferential intervals. Deployment is selected by target-domain P1 only; a different rule would select different trajectory points. P2 is not attribute-isolating, and neither P3 nor CAPP establishes spatial grounding, localization, or part binding. For CUB/NABirds and FineView, source descriptions are LLM-processed from public species resources, whereas ImageNet descriptions and attributes are LLM-generated (Appendix B.2); residual source, template, or generation effects cannot be fully excluded.

Expanded interpretive controls. Three controls constrain this interpretation (full numbers in Appendix H). In the representative DFN5B/CUB setting, WiSE interpolation (α=0.5) partially recovers held-out P3 for N and U while retaining most target P1/P3, so specialization need not be irreversible; this is not a general WiSE result, and SSR/SSR+ND are not tested because their structured components lack direct pretrained counterparts. The direct name-access ablation (§4.2) shows that restricting direct name access changes P3/CAPP and attribute-route geometry while ordinary P2 remains name-dominated, a routing effect distinct from the late-specialization analyses in §4.4. A target-side hard-negative objective improves target P3 without a corresponding held-out advantage at matched later checkpoints, showing that stronger in-domain semantic discrimination need not imply stronger transferable accessibility at the same stage of training (Appendix H.2).

J.2 Main-item to supplement map

Main item	Supplement table(s)	Population	Checkpoint rule	Evidence type
Fig. 2b	Tables S11–S15	ImageNet target, same-ontology shifts, cross-ontology; six backbones	trajectory peak	six-backbone descriptive
Fig. 2a	Table S14	CUB target, NAB-348 held-out; DFN5B	peak, own-P1 deploy	single run per method
Table 1	Table S19	ImageNet val; six backbones	own-P1 deploy	six-backbone descriptive
Fig. 3a	Table S20	CUB target, NAB-348; DFN5B	own-P1 deploy	single run, 1000-replicate class bootstrap
Fig. 3b	Table S21	3 training domains $\times$ six backbones (18 pairs)	trajectory peak	six-backbone descriptive
Table 2	Tables S22–S24	CUB/NAB target; six backbones	own-P1 deploy	six-backbone descriptive
Fig. 4a	Table S14	DFN5B CUB-trained, CUB-200 and NAB-348	peak, deploy	single run per method
Fig. 4b	Tables S30, S32, S33, S11–S13	six backbones; target and cross-ontology, by method	categorical peak vs. deploy	six-backbone descriptive
Fig. 4c	Table S31, S13	six backbones; cross-ontology cells with $t_{P3} < t_{\text{deploy}}$	peak $-$ deploy P3 (pp)	six-backbone descriptive
Fig. S10	Table S29	DFN5B CUB-trained, CUB-200 and NAB-348	P3 / CAPP peak, deploy	single run per method
Fig. 5a	Tables S34–S37	N/U/SSR grid; joint image+text backbone movement (encoder-separated audit in Tables S35–S36)	own P3 peak / P1 deploy	descriptive association
§G.4 (no main panel)	Tables S39, S40	DFN5B SSR+ND, NAB-348	epoch 20 / 60	single run
§G.5 (no panel)	Table S41	three additional SSR+ND runs (MetaCLIP H/14, DFN2B L/14, MetaCLIP L/14)	own P3 peak / deploy	independent runs, class bootstrap
Fig. S11	§G.2	same 78 held-out N/U/SSR cells as Fig. 5a, held-out P1 at P3-peak vs deploy	same intervals	descriptive association
§G.3 (no panel)	Table S38	SSR+ND DFN5B CUB$\rightarrow$NAB-348, P3 and P1, peak-ranked wrong classes	peak / deploy	single run, class bootstrap
Fig. 5b	Tables S44–S45	SSR+ND (per-class values); text-frozen U control in the table, NAB-348 P3	peak / deploy	single runs, class bootstrap
Fig. S12a	Table S43	SSR+ND; text-frozen U	PT / peak / deploy	single runs
Fig. S12b	Table S42	SSR+ND, NAB-348	four encoder states	single run
Discussion: WiSE	Table S48	DFN5B CUB-trained, N and U	interpolation α	single seed
Discussion: hard negative	Fig. S13, Table S49 (§H.2)	DFN5B CUB-trained; CUB target, NAB-348 strict	matched epochs (own-deployment comparison in §H.2)	single seed, single λ
Discussion: name access	Tables S25–S28, Fig. S4	DFN5B CUB-trained; CUB target, NAB-348 strict	own-P1 deploy (P3 and CAPP trajectories)	single run per variant
§F.1 (P1/P3 atlas)	Figs. S5–S9	ImageNet-trained (DFN/MetaCLIP, OpenAI); CUB-/NABirds-trained (DFN/MetaCLIP, OpenAI); target, full and strict held-out kept separate	own-P1 deploy marker; no separate P1 optimum	single run per method and backbone
P3 vs. CAPP change	Fig. S3, Tables S22–S24	DFN/MetaCLIP and OpenAI; target and strict held-out	own-P1 deploy	descriptive, no coefficient

Table S51: Main-paper items mapped to supplement tables, evaluation populations, checkpoint rules, and evidence type. Evidence types are distinct: single training runs, six heterogeneous backbones (descriptive), class-level bootstrap resampling of the evaluation set, and archived historical multi-seed checks (Table S7); they must not be conflated.

J.3 Experiment provenance and run naming

SSR is the structured NAME/ATTRIBUTE representation with routed causal attention and Structured Text Fusion defined above (CUB/NAB include a shared CONCEPT segment; ImageNet does not), and SSR+ND is SSR with NAME-summary dropout ($p = 0.5$) at training-time fusion. For DFN5B/CUB the canonical SSR+ND run differs from the SSR run only by the added name-drop and is distinct from an earlier historical name-drop calibration run. The historical run is reported only in the supplemental probability-calibration analysis (§C.8) and is not used in the canonical main

trajectory results. Historical multi-seed runs used earlier SSR routing variants that are not used for the main results and are not the canonical SSR.

J.4 Limitations of replication and seed coverage

All main trajectory results come from one training run per backbone, domain, and method (seed 0); variability across the six backbones is descriptive heterogeneity, not a confidence interval. The mechanism, WiSE, hard-negative, name-access, and text-frozen analyses are single representative runs. Multi-seed evidence exists only for historical DFN5B configurations and only for in-domain retrieval. Bootstrap intervals quantify evaluation-set sampling uncertainty only. Unresolved provenance items: the CAPP evaluator source that averages cells into the class-slot-macro score; the NAB-348 derivation script; OpenAI ImageNet parameter files for steps 1–5; GPU type for the untyped allocations and exact per-run software versions.