WHEN BETTER ROUTING HIDES A WORSE PROMPT UPDATE: CLASSIFIER-CONDITIONED EVALUATION IN DOMAIN-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt-only changes arise when prompt-based continual learners select or compose prompts under a shared readout, or when a selector or prompt bank is updated while a deployed classifier is retained. Joint accuracy correctly evaluates deployment of a complete prompt–classifier branch, but it does not reveal whether the prompt-side change helps under the retained classifier. We show a concrete decision reversal: on S-iPrompts DomainNet, a Gaussian router that uses no test-domain labels improves domain identification and joint routed accuracy by 1.28 percentage points (pp), yet reusing only its prompt assignments with the original classifier assignments lowers accuracy by 0.23 pp. Across four configurations on Office-Home and DomainNet, the same prompt correction can reverse sign solely with the classifier that reads it; this occurs in S-iPrompts and an independently implemented DualPrompt bank with added per-domain heads, whereas the correction helps under both classifiers in CP-Prompt. On DomainNet, matched classifier refits also change this dependence in all three per-domain-head configurations without improving routed accuracy over their matched baselines. We therefore introduce *intervention-matched evaluation*: score prompt-only changes under the classifier that will read them, joint changes by joint accuracy, and training changes against a matched refit, reporting paired repairs and harms. This prevents evidence for one deployment action from approving another.

1 INTRODUCTION

Models encounter new data over time, but storing and repeatedly retraining on all past data may be impractical. Updating on new data can also erode earlier capabilities, a problem known as catastrophic forgetting (Kirkpatrick et al., 2017; Parisi et al., 2019; De Lange et al., 2022). Continual learning aims to adapt through this sequence while retaining those capabilities. In *domain-incremental learning* (DIL), one such setting, classes remain the same while the input distribution changes, as when recognising chairs and cars across photographs and drawings. Test inputs come without domain identity; the model must recognise examples from the encountered domains (van de Ven et al., 2022; Wang et al., 2022a; Feng et al., 2024).

Prompt tuning offers a parameter-efficient way to adapt a large pretrained model by learning continuous inputs or internal prefixes while keeping the backbone fixed (Lester et al., 2021; Li & Liang, 2021; Jia et al., 2022). Here, *prompts* are small sets of learned vectors, not natural-language instructions. The image-only S-Prompts variant, S-iPrompts, learns separate prompts for successive domains and preserves earlier prompts, thereby limiting direct overwriting of prior adaptations (Wang et al., 2022a).

Prompt banks create an inference-time choice, even when improving that choice is not the method's primary goal. In L2P, native DualPrompt, and CODA-Prompt, prompt selection or composition feeds a shared classifier, so a prompt-side change retains that classifier by construction (Wang et al., 2022c;b; Smith et al., 2023). S-Prompts and CP-Prompt instead retain domain-specific prompts and classifiers; a *router* chooses a pair for each test input without its domain label (Wang et al., 2022a; Feng et al., 2024). Changing prompt assignments or the bank while retaining deployed classifier assignments is a prompt-only intervention; deploying the selected pair is a joint intervention.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

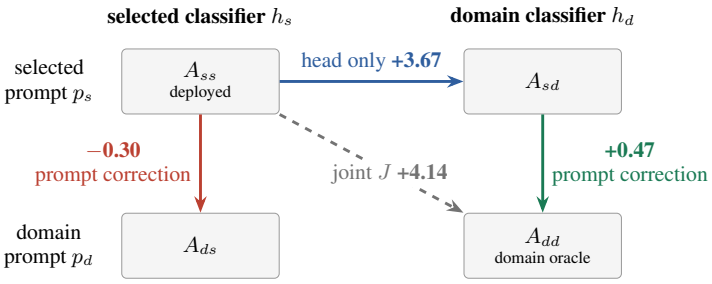

Figure 1: **The same prompt correction in two classifier contexts.** Each cell scores the same images with the prompt (row) and retained classifier (column) set to the selected (s) or true-domain (d) assignment; subscripts denote prompt then classifier. Arrows give S-iPrompts DomainNet accuracy changes in pp (three-seed means; Table 2). Holding a classifier fixed means retaining its assignment per image, not using one classifier for the entire dataset.

Retaining the readout also matters beyond shared-readout systems. It can be an explicit constraint when replacing only a prompt selector or module; if prompt features change, unavailable past training data may limit refitting earlier classifiers. Even when the full pair will be deployed, fixing the original classifier assignments measures the prompt-only effect under that readout.

Ground-truth domain identity offers a diagnostic: replace selected parameters with the true-domain branch, as in SOYO's oracle-routing evaluation (Wang et al., 2025b). The resulting joint comparison estimates the full-branch update. We test how classifier choice and training exposure affect the corresponding prompt-side action.

This distinction produces a concrete deployment-decision reversal. In S-iPrompts on DomainNet, a Gaussian router whose protocol was locked before test scoring and which uses no test-domain labels improves domain identification and gains 1.277 percentage points (pp) when it selects both prompts and heads. Retaining the heads selected by the original router and applying only its prompt assignments instead loses 0.225 pp (Section 4.1). The joint metric correctly accepts the full prompt–head update; using it to approve the prompt-only deployment would choose a harmful action.

The same banks strongly depend on their prompts: removing them costs 23.76 pp. Reliance on prompts, the benefit of correcting their identity, and the gain from changing a prompt–classifier pair therefore answer different design questions. Figure 1 shows how the retained classifier changes the answer even for true-domain prompt correction.

We ask two empirical questions. First, can the same prompt correction change sign solely because the retained classifier changes? Second, can changing classifier exposure move this dependence without improving the predictor? Frozen-bank substitutions answer the first; matched refits test the second. We then translate the results into evaluation rules for prompt-only, joint, and training interventions.

Our contributions follow this chain from diagnosis to evaluation:

- **A component-level diagnosis.** Across four evaluated configurations and two benchmarks, paired substitutions separate reliance on prompts from the value of correcting them and, where per-domain heads (PDH) exist, from joint correction. On DomainNet, the retained classifier reverses the sign of prompt correction in S-iPrompts and DualPrompt + PDH, an independently implemented DualPrompt bank augmented with PDH. In S-iPrompts, it also reverses the effect of a router requiring no test-domain labels that improves both routing and routed accuracy (Section 4.1).
- **A controlled test of classifier exposure.** Matched refits show a signed interaction response on DomainNet in all three per-domain-head configurations; in two, data-budget controls separate it from reduced own-prompt data. Further seeds and a reversed order reproduce the S-iPrompts response (Section 4.2).
- **A protocol for component-level design decisions.** Specify the changed component, retained classifier, reference predictor, and scored population; then report paired accuracy,

repairs, and harms. This separates decisions to change a prompt, a pair, or classifier training, preventing evidence for one action from approving another (Sections 4.3–5).

Related work already treats task inference and prompt–classifier compatibility as important. CPrompt trains for prompt–classifier consistency (Gao et al., 2024); CPS-Prompt separates prompt and linear-head fitting (Lim et al., 2026); HiDe-Prompt shows strong classification despite poor task inference under particular representations (Wang et al., 2023). These results motivate the question, but joint selection metrics do not estimate a prompt-only change under a retained classifier. Our fixed-bank crossings measure that intervention directly. This connects to model stitching, which tests compatibility across network parts (Lenc & Vedaldi, 2015; Bansal et al., 2021), and backward-compatible training, which preserves compatibility with previously computed representations (Shen et al., 2020).

Together, our findings yield a practical rule for banked continual learners: judge a router's prompt-side value by whether its prompt choices improve predictions under the classifier that will read them, not by domain-identification or joint routed accuracy alone.

2 Evaluating a prompt change under its classifier

The unit of comparison. The evaluated counterfactual must match the proposed deployment. With a shared readout, prompt-only value is the direct accuracy change under that readout. When prompt and head assignments can vary separately, compare the selected and true-domain prompts under each of two retained classifiers: the selected classifier and the true-domain classifier (Figure 1). Throughout a substitution, the image, trained parameters, and original routing assignment stay fixed, so the four cells evaluate a completed model rather than a retrained one.

Setting and notation. In DIL, domains share a label space (van de Ven et al., 2022), so every prompt–classifier pairing predicts over the same classes. Let S contain examples (x, y, d), with true domain d, and let $s(x)$ be the original router assignment. A *bank* is a final trained checkpoint containing the available prompts, classifiers, and router. We use *head* for a per-domain classifier and *readout* for any prediction rule, including CPrompt's combined classifier. For prompt assignment π and head assignment h, define

$$A_h(\pi; S) = \frac{1}{|S|} \sum_{(x,y,d)\in S} \mathbf{1}[f(x; \pi(x), h(x)) = y]. \tag{1}$$

The shorthand $A_{ab} = A_{h_b}(a; S)$ uses $a, b \in \{s, d\}$. Accuracy differences are reported in pp, i.e. 100 times the corresponding difference of proportions. Unlike disjoint-label tasks with disjoint output heads, this shared-label setting permits crossed pairings without excluding the correct class by construction.

What the four cells identify. The prompt-only action is $A_{ds} - A_{ss}$, the head-only action is $A_{sd} - A_{ss}$, and the joint domain-correction gain is $J = A_{dd} - A_{ss}$. Their interaction is

$$I_{PH} = (A_{dd} - A_{sd}) - (A_{ds} - A_{ss}) = A_{dd} - A_{ds} - A_{sd} + A_{ss}. \tag{2}$$

Thus I_{PH} is the difference between two prompt-correction effects, not a separate intervention. If it is nonzero, the effect of correcting the prompt depends on the head. Algebraically, J is the sum of the prompt-only action, head-only action, and interaction. Ratios of an action to J describe these measured contrasts; they do not assign causal shares of knowledge to components. The joint contrast uses only the two diagonal cells, whereas the component effects also require the off-diagonal cells.

Dependence and correction are different interventions. Removing a trained prompt asks whether the deployed model relies on it. Replacing it asks whether another prompt is better under the retained readout. We keep their references explicit:

$$D_h(p; S) = A_h(p; S) - A_h(\varnothing; S), \quad p \in \{s, d\}, \qquad \text{removal cost,} \tag{3}$$

$$C_h(S) = A_h(d; S) - A_h(s; S), \qquad \text{domain correction gain.} \tag{4}$$

Here $\varnothing$ removes prompts without refitting. Removal characterises reliance of this trained predictor; it does not compare architectures trained with and without prompts. Unless stated otherwise, main-text removal uses the selected prompt and deployed readout as its reference; Appendix B also reports the domain-prompt/domain-head reference and explains why its removal cost differs.

Separate the events behind a net gain. Let $R_u(S)$ and $H_u(S)$ be the fractions of images repaired or harmed by intervention u, relative to the selected prompt under the same head. Let $\mathcal{M} = \{x : s(x) \neq d(x)\}$ be the original domain-mismatch set. Then

$$A_h(u; S) - A_h(s; S) = R_u(S) - H_u(S), \qquad C_h(S) = \frac{|\mathcal{M}|}{|S|} C_h(\mathcal{M}). \tag{5}$$

An unchanged net accuracy can hide both repairs and harms; studies of model updates call the latter negative flips (Yan et al., 2021). A change between two wrong classes is neither. We compute all contrasts within each seed before averaging, and distinguish full-test effects from effects conditional on $\mathcal{M}$.

Table 1: **Evaluated configurations and their roles.** PDH denotes the per-domain linear heads added to the released DualPrompt model; this is an evaluation adaptation, not the native DualPrompt classifier.

Configuration	Prompt/readout structure	Role in the evidence
S-iPrompts	Domain input prompt; per-domain linear head	Primary fixed-bank and router tests
CP-Prompt (CLIP)	Common/personalised visual prompts; text-generated classifier	Positive prompt-only boundary
DualPrompt+PDH	General/expert prompts; added per-domain linear heads	Independent-code controlled check with added heads
CPrompt (combined)	Domain prompt bundle; combined log-sum-exp readout	Boundary without a swappable head

3 STUDY DESIGN

Four configurations play distinct evidential roles. We evaluate twenty-one checkpoints on Office-Home (OH) and DomainNet (DN), with three training seeds per configuration and benchmark (DualPrompt+PDH: DN only). Table 1 states which readout each configuration exposes and why it is included; Appendix A records checkpoint provenance.

S-iPrompts denotes the image-only variant of S-Prompts (Wang et al., 2022a; Dosovitskiy et al., 2021). CP-Prompt uses CLIP image and text encoders (Feng et al., 2024; Radford et al., 2021). DualPrompt+PDH starts from an independent DualPrompt code base; the released model has one shared classifier, so we add one linear head per domain and train it with that domain's expert prompt (Wang et al., 2022b). Appendix H records all changes. CPrompt is our shared-label DIL port (Gao et al., 2024). The four-cell and exposure experiments use the fifteen per-domain-head banks. Removal covers all twenty-one; exhaustive prompt substitutions (Section 4.3) cover the original eighteen. Appendix A gives implementation details.

A controlled change in classifier exposure. To test whether the features seen during classifier training change the prompt-correction effect, we refit each configuration-specific head from its already-trained state in two matched arms. *Own* uses its original partner prompt throughout. *Mixed* uses that prompt on 50% of fit examples and uniformly assigns the other prompts to the remainder. Prompts, backbone, assignments, image sequence, labels, optimiser, and update budget are fixed across arms. For CP-Prompt, refitting updates its text context rather than substituting a linear head. Both arms use the same configuration selected by their mean validation accuracy from a grid locked before test scoring. Own-half controls separate foreign-prompt exposure from reduced own-prompt data.

Before test evaluation, we designated the S-iPrompts DN interaction and harm contrasts as primary and CP-Prompt as the secondary configuration, and fixed the selection rule and 0.5 pp accuracy margin (Appendix F). DualPrompt+PDH's gates and criteria were fixed before its training began (Appendix H). This post-training refit is a diagnostic intervention inspired by prompt-consistency training (Gao et al., 2024) and post-hoc classifier alignment (Zhang et al., 2023), not a proposed learner.

Table 2: **Prompt–head substitution at fixed checkpoints, all three per-domain-head configurations.** Full-test accuracy changes in pp, mean ± sample SD over three seeds. Subscripts denote prompt then head; s is selected and d is domain-matched. The last row changes only the head context relative to the first row. CP-Prompt uses its trained-bank arm from the twelve-bank consolidation in Appendix F; DualPrompt+PDH is the evaluation adaptation defined in Table 1.

	S-iPrompts		CP-Prompt		DualPrompt+PDH
Intervention or contrast	Office-Home	DomainNet	Office-Home	DomainNet	DomainNet
Prompt only: $A_{ds} - A_{ss}$	-0.301 ± 0.289	-0.297 ± 0.064	$+0.685 \pm 0.093$	$+0.902 \pm 0.106$	-0.114 ± 0.051
Head only: $A_{sd} - A_{ss}$	$+2.397 \pm 0.112$	$+3.669 \pm 0.039$	$+0.643 \pm 0.405$	$+1.138 \pm 0.042$	$+3.845 \pm 0.058$
Joint: $A_{dd} - A_{ss}$	$+2.438 \pm 0.279$	$+4.141 \pm 0.038$	$+1.235 \pm 0.072$	$+2.194 \pm 0.031$	$+4.061 \pm 0.027$
Interaction: I_{PH}	$+0.342 \pm 0.124$	$+0.769 \pm 0.117$	-0.093 ± 0.316	$+0.154 \pm 0.094$	$+0.330 \pm 0.028$
Prompt with domain head: $A_{dd} - A_{sd}$	$+0.041 \pm 0.171$	$+0.472 \pm 0.076$	$+0.591 \pm 0.408$	$+1.056 \pm 0.067$	$+0.216 \pm 0.070$

Populations and replication. OH has four domains and 65 classes; our fixed within-domain split supplies 3,213 test examples (Venkateswara et al., 2017). DN has six domains and 345 classes; we use its 176,743 official test examples and reserve part of the training data for validation (Peng et al., 2019). We report final-stage, example-weighted accuracies over seen domains. Primary comparisons use one fixed order per benchmark and three-seed means with sample SDs. Five additional S-iPrompts DN checkpoints test seed and order replication, yielding five canonical-order seeds and three reversed-order seeds (Appendix G).

For all main prompt-choice and removal results, we recompute predictions for every test image and verify the resulting aggregate metrics. Each paired substitution uses the same images and recorded prompt assignments; repairs and harms directly count images that change from incorrect to correct or from correct to incorrect, respectively. Image-level tests and paired bootstraps compare alternatives within each fitted bank, whereas seed SDs measure variation across separate training runs. Prompt substitutions change choices only at inference time, while exposure experiments refit only the classifiers.

4 RESULTS

We first establish the classifier-conditioned prompt effect and delimit where it holds (Section 4.1). We then test classifier exposure as one factor that changes this effect (Section 4.2). Finally, we decompose a small net gain into repair and harm events and ask whether other prompt assignments carry value (Section 4.3).

4.1 THE CLASSIFIER CAN CHANGE WHETHER PROMPT CORRECTION HELPS

One prompt correction, two outcomes. The central comparison is the same prompt correction under two retained heads. In S-iPrompts on DN, it changes accuracy by -0.297 ± 0.064 pp under the selected head and $+0.472 \pm 0.076$ pp under the domain head (Table 2); the interaction is $+0.769 \pm 0.117$ pp. DualPrompt+PDH, built on an independent code base and augmented with PDH for this evaluation, shows the same reversal: -0.114 ± 0.051 pp under the selected head and $+0.216 \pm 0.070$ pp under the domain head, with these signs in every seed (exact McNemar $p < 0.05$). In both configurations, the prompt's true-domain identity is the same in the two comparisons, yet its predictive effect changes sign. This is why the classifier belongs in the definition of the evaluated prompt change. Head dependence may seem expected, since each head is trained with its own prompt. The reversal is not: CP-Prompt's heads are also trained with their own prompts, yet prompt correction helps under both.

A better joint router can be a worse prompt-only update. Under a protocol locked before test scoring, the Gaussian router raises S-iPrompts DN routing accuracy from 80.43% to 83.30% and, when selecting both components, routed accuracy by 1.277 ± 0.045 pp (Appendix D). Applied to prompts alone, its assignments change accuracy by -0.225 ± 0.029 pp under the originally selected heads and by $+0.371 \pm 0.074$ pp under the heads selected by that router, with the same signs in all five canonical-order banks. The Gaussian router therefore passes the joint-deployment test and fails the prompt-only test under the original heads. The router uses no test-domain labels; it fits domain

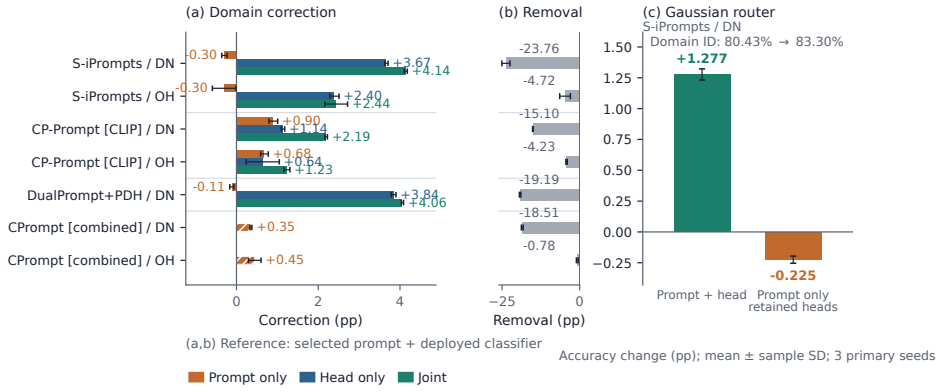

Figure 2: **A better joint router can be a worse prompt-only update.** (a) Domain-prompt, domain-head, and joint correction under the original selected-prompt reference; CPrompt exposes only prompt correction under its combined readout. (b) Prompt removal under the same reference (separate scale). (c) The S-iPrompts DomainNet Gaussian router improves domain identification and joint accuracy, but its prompt assignments reduce accuracy under the original head assignments. Bars are three-seed means; whiskers show sample SD. DN/OH: DomainNet/Office-Home.

distributions only on training features. Two alternate routers that reduce domain identification (L_2, cosine; Figure 7) show the same classifier-dependent pattern. Under CPrompt's combined readout, by contrast, the Gaussian prompt assignments raise accuracy.

Prompt reliance does not identify correction value. S-iPrompts strongly relies on its prompts: removing them costs 23.76 ± 1.30 pp on DN, and even a head refitted on prompt-free features remains 10.10 ± 1.13 pp below the matched prompted refit (Table 18). Yet replacing the selected prompt by the true-domain prompt changes accuracy by only -0.297 ± 0.064 pp under the retained head. By contrast, the head-only and joint actions gain $+3.669 \pm 0.039$ and $+4.141 \pm 0.038$ pp (Figure 2); the former reaches 89% of the latter while retaining the originally selected prompt (95% in DualPrompt+PDH). These are ratios of measured actions, not causal shares. The joint gain therefore does not measure prompt-only correction. On OH, the selected-head effect is also negative in every seed, and its head-only action reaches 98% of the joint gain; its domain-head effect, $+0.041 \pm 0.171$ pp, is not separated from zero, so the reversal is established on DomainNet.

Prompt-only effects differ across configurations. CP-Prompt's selected-head prompt correction is positive in every seed on both benchmarks: $+0.902 \pm 0.106$ pp on DN and $+0.685 \pm 0.093$ pp on OH. On DN its head-only action is $+1.138 \pm 0.042$ pp against a joint gain of $+2.194 \pm 0.031$ pp. The head-only action reaches 52% of the joint gain; the two single-component actions have similar magnitudes, unlike S-iPrompts. All three joint gains are positive, yet the prompt-only action is negative in S-iPrompts and DualPrompt+PDH and positive in CP-Prompt. A positive joint gain therefore does not determine the sign of the prompt-only effect; the result also does not imply that prompt correction is generally harmful.

CPrompt supplies a different boundary: its combined readout permits prompt removal and correction, but no head-only swap. On DN, removal costs 18.51 ± 0.29 pp while domain-prompt correction gains $+0.353 \pm 0.024$ pp. Across all four configurations, prompt removal costs 15–24 pp on DN (Figure 2b), whereas prompt correction under the deployed readout ranges from approximately -0.3 to $+0.9$ pp. Mismatch masses of 16–23% thus coexist with different correction values.

The selected-head loss survives domain weighting, further seeds and orders. All three checks use S-iPrompts. Quickdraw and real account for 59% of the DN test images, so we also weight the six domains equally (Table 16). Prompt-only correction remains negative in every seed (-0.201 ± 0.044 pp); its domain-level mean is negative in four domains, positive in painting, and near zero in quickdraw. The mean head-only effect exceeds the mean prompt-only effect in every domain, and its domain-mean gain reaches 85% of the joint gain (89% pooled).

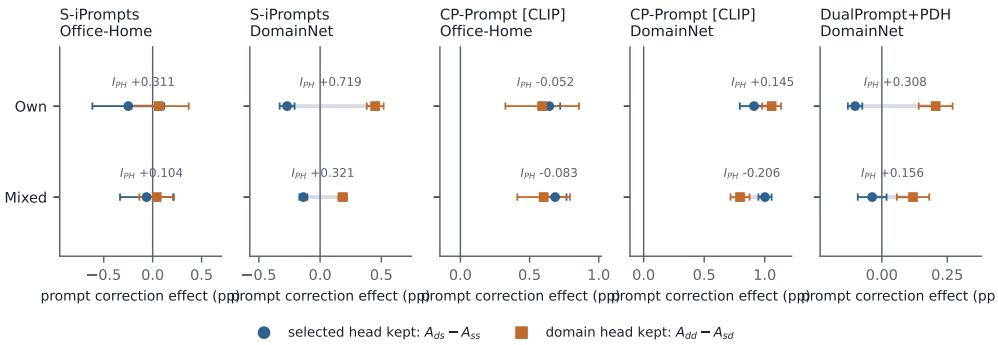

Figure 3: **Classifier refitting changes the effect of the same prompt correction.** Both refits retain the prompts and assignments. Each row shows prompt correction under the selected head ($A_{ds} - A_{ss}$) and domain head ($A_{dd} - A_{sd}$); their separation is I_{PH}. Whiskers: sample SD over three seeds. On DN, the two effects move closer together in S-iPrompts and DualPrompt+PDH, and CP-Prompt's interaction crosses zero. Table 3 reports the accuracy costs.

The selected-head prompt loss also persists across five canonical-order seeds and three reversed-order banks (-0.306 ± 0.076 and -0.277 ± 0.007 pp). A between-bank comparison with shared routing assignments (Appendix E) also finds classifier-dependent value: on DN, joint domain correction gains 0.137 pp with shared heads and 2.819 pp with PDH.

4.2 Classifier exposure moves component effects, not necessarily accuracy

The frozen-bank experiment locates the dependence on the retained classifier; whether classifier training moves it is the next question. The natural candidate is training exposure: a head fitted to one prompt's features may respond differently to another prompt. The primary contrasts, fixed before test scoring, are the signed changes in the interaction and in prompt-switch harm between matched Mixed and Own refits. A change in these contrasts tests exposure as a factor; an accuracy improvement is a separate outcome.

Mixed exposure narrows the S-iPrompts interaction. On DN, the interaction falls from 0.719 ± 0.112 pp in Own to 0.321 ± 0.059 pp in Mixed, a paired change of -0.398 ± 0.053 pp that is negative in every seed. Prompt correction moves from -0.270 to -0.137 pp under the selected head and from $+0.450$ to $+0.185$ pp under the domain head (Figure 3). The two effects move closer together, but correcting the prompt still has opposite signs under the two heads. Two further canonical seeds and the reversed-order banks reproduce the I_{PH} decrease (five-bank and reversed-bank means -0.395 ± 0.038 and -0.397 ± 0.107 pp; Appendix G).

The signed exposure response repeats across configurations, but it is not an accuracy improvement. DualPrompt+PDH's interaction falls from 0.308 pp in Own to 0.156 pp in Mixed, a paired change of -0.152 ± 0.029 pp that is negative in every seed. Harm changes by -0.068 ± 0.012 pp and repair by -0.003 ± 0.017 pp, while routed accuracy falls by 0.32 pp, inside the 0.5 pp margin used for the other configurations. CP-Prompt's signed interaction also decreases in every DN seed, by -0.351 ± 0.106 pp, but crosses zero from $+0.145$ to -0.206 pp, so its magnitude grows in two of three seeds. Its routed and domain-pair accuracies fall by 1.023 ± 0.073 and 0.877 ± 0.074 pp, beyond the margin fixed before test scoring. The signed response is consistent; the performance outcome is not an improvement.

The diagonal joint contrast hides off-diagonal movement. The joint correction gain changes by -0.001 ± 0.041 pp in S-iPrompts, $+0.029 \pm 0.014$ pp in DualPrompt+PDH, and $+0.146 \pm 0.068$ pp in CP-Prompt. In S-iPrompts, both diagonal accuracies fall by 0.108 pp on average, while the prompt-only action, head-only action, and interaction change by $+0.133$, $+0.264$, and -0.398 pp, nearly cancelling in J. Measuring the off-diagonal cells reveals the altered component effects; the joint contrast alone cannot identify them.

Table 3: **An interaction change must be read alongside predictive performance.** Mixed minus matched Own refits, full-test pp; paired three-seed mean ± sample SD. Repair and harm refer to prompt correction under the selected head relative to each arm's own baseline. CPrompt has no paired head to refit and no exposure arm.

Family	Data	ΔI_{PH}	Δ harm	Δ repair	ΔA_{ss}	ΔA_{dd}
S-iPrompts	OH	-0.207 ± 0.406	-0.104 ± 0.140	$+0.083 \pm 0.090$	-0.114 ± 0.177	-0.041 ± 0.234
S-iPrompts	DN	-0.398 ± 0.053	-0.187 ± 0.015	-0.054 ± 0.022	-0.108 ± 0.046	-0.108 ± 0.023
CP-Prompt	OH	-0.031 ± 0.194	-0.010 ± 0.065	$+0.031 \pm 0.054$	-0.114 ± 0.252	-0.062 ± 0.300
CP-Prompt	DN	-0.351 ± 0.106	-0.089 ± 0.036	$+0.002 \pm 0.054$	-1.023 ± 0.073	-0.877 ± 0.074
DualPrompt+PDH	DN	-0.152 ± 0.029	-0.068 ± 0.012	-0.003 ± 0.017	-0.319 ± 0.006	-0.290 ± 0.016

Controls compare foreign-feature exposure with continued fitting and data dilution. Own is needed because further fitting alone changes the classifier: in S-iPrompts DN, it raises routed accuracy by 0.617 ± 0.202 pp but moves I_{PH} by only -0.049 ± 0.022 pp (Appendix F). Comparing Mixed with Own therefore removes the effect of continued fitting.

Mixed also changes the number of own-prompt examples. Own-half controls retain Mixed's own-prompt slots while omitting foreign features, evaluated at two budgets. Across the DN control conditions in S-iPrompts and CP-Prompt, mean interaction changes have magnitude at most 0.012 pp, compared with the 0.398 and 0.351 pp shifts under Mixed (Appendix F, Table 17). These comparisons support an effect of foreign-feature exposure rather than reduced own-prompt data alone.

Read the diagnostic response alongside accuracy. In S-iPrompts DN, prompt-correction harm falls by 0.187 ± 0.015 pp, but repair also falls, by 0.054 ± 0.022 pp, and routed accuracy declines by 0.108 ± 0.046 pp (Table 3).

A lower baseline accuracy can mechanically reduce measured harm by leaving fewer correct predictions to damage. We therefore partition images by their baseline correctness in both fitted arms (Table 21). Images correct in both contribute -0.236 ± 0.020 pp to the S-iPrompts DN harm contrast; the other partitions partly offset this to yield -0.187 pp overall. CP-Prompt also has a negative common-correct contribution in every DN seed. The partition shows that lost baseline correctness does not fully explain the harm decrease.

On DN, the image-bootstrap intervals in Table 20 exclude zero for every listed contrast in all three per-domain-head configurations. On OH, where one pp is about 32 test images, the intervals include zero, so the OH exposure results are inconclusive on their own.

4.3 A SMALL CORRECTION GAIN CAN CONCEAL DIFFERENT PREDICTION CHANGES

Once the intended classifier is fixed, a small net gain still needs interpretation. It may reflect unchanged predictions, repairs offset by harms, or changes among wrong classes. In S-iPrompts DN under the selected head, domain-prompt correction changes predictions on 28.05% of the original mismatch set (integer counts and estimator in Appendix B). It repairs $3.584 \pm 0.084\%$ and harms $5.101 \pm 0.261\%$ of that set, yielding a conditional net of -1.517 ± 0.337 pp. Another 19.37% of the set changes from one wrong prediction to another. Thus complete prediction invariance is ruled out for this intervention, and repairs offset by harms account for only part of its prediction changes.

Appendix C tests two further boundaries. Under a fixed domain head, the original S-iPrompts and CP-Prompt assignments outperform their frequency-preserving permutations within every DN mismatch set. A label-informed choice among all S-iPrompts prompts gives a 6.654 pp full-test gain, but 4.812 pp lies outside that set and the gap can reflect generic error diversity. These analyses reveal structure in the finite bank; neither supplies a deployable selector.

5 IMPLICATIONS: EVALUATE THE INTERVENTION THAT WILL BE DEPLOYED

When does correcting a prompt help? In the evaluated DomainNet banks, domain-prompt correction helps under the domain classifier in every seed of all three per-domain-head configurations.

Table 4: **Choose the measurement after specifying the proposed change.** Use the same scored images and name the reference predictor. Report accuracy and paired repair/harm counts; a favourable surrogate diagnostic alone does not establish an improvement.

Proposed change	Keep fixed	Evidence needed
Prompt only	Intended classifier per image	Prompt-only accuracy change, repairs and harms
Prompt and head	Bank and evaluation population	Joint gain; four cells if independently selectable
Combined-readout selection	Native combined readout	Prompt changes under that readout
Classifier exposure	Prompts, assignments, fit budget	Matched-refit contrasts, signed and absolute interaction, accuracy cost

Under the selected classifier, its sign is negative in S-iPrompts and DualPrompt+PDH but positive in CP-Prompt. Mixed-prompt classifier exposure narrows this difference in S-iPrompts and DualPrompt+PDH (Section 4.2). The retained classifier and its exposure, not domain identity alone, shape correction value.

Match the measurement to the proposed deployment. Score a prompt or selector change under the retained readout, a joint prompt–head change by joint accuracy, and a training change against a matched refit with actual accuracy. When the components are independently selectable, the four cells explain the associated single-component actions (Table 4).

A prompt-selector decision. For a banked system with per-domain classifiers, a joint domain-oracle gain also changes the classifier assignment and therefore does not estimate a prompt-only selector update. Apply the proposed prompts under the originally selected heads, then count repairs and harms. For the Gaussian router in Section 4.1, routing accuracy and joint gain improve while retained-head prompt accuracy falls. If both components will change, joint accuracy is the right outcome.

A classifier-training decision. For training intended to improve prompt–classifier compatibility, the four cells measure dependence, but evaluation also needs the original predictor, relevant counterfactual, and an equal-budget refit. Our results show why: lower harm can accompany fewer repairs, and a consistent signed interaction shift can accompany larger absolute interaction and lower accuracy. A favourable diagnostic alone cannot approve the training change.

A small domain-prompt correction does not make a shared prompt sufficient: tied-prompt banks lose 4.67 pp before classifier refitting reduces the mean deficit (Appendix F.1).

Scope of the conclusions. Our claims concern the trained banks evaluated here: four ViT-based configurations, one primary domain order, and three seeds per setting, with extended seed and order replication for S-iPrompts on DomainNet. Appendices I and J state further boundaries and relate the protocol to task-inference and expert-selection work.

6 Conclusion

We showed that joint evidence can approve a prompt update that hurts under the classifier that will read it. On S-iPrompts DomainNet, a label-free Gaussian router improves domain identification and joint accuracy, yet its prompt assignments lower accuracy under the original classifier assignments; on DomainNet, the same true-domain prompt correction reverses sign across classifiers in S-iPrompts and DualPrompt+PDH, and matched classifier refits move this dependence without improving accuracy. Evaluating each change under the predictor that will use it (prompts under their reading classifier, prompt–classifier pairs by joint accuracy, and training changes against matched refits, with repairs and harms) separates evidence for a joint update from evidence for its prompt alone.

Reproducibility statement

Appendices A–H specify the datasets, interventions, aggregation, sample manifests, preprocessing, and reconstruction checks. The accompanying source package, available at `https://anonymous.4open.science/r/prompt-head-decomposition-F1C1/README.md`, contains the per-seed summaries, the code that regenerates the reported figures, and the aggregation code behind the reported tables.

Use of large language models

Large language models assisted with drafting, code preparation, and consistency checks. Numerical claims were checked against the archived experiment outputs and executable aggregation code; this assistance did not create additional empirical results. The authors take full responsibility for all content.

References

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Rafael MO Cruz, Robert Sabourin, George DC Cavalcanti, and Tsang Ing Ren. META-DES: A dynamic ensemble selection framework using meta-learning. *Pattern recognition*, 48(5):1925–1935, 2015.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

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

Yu Feng, Zhen Tian, Yifan Zhu, Zongfu Han, Haoran Luo, Guangwei Zhang, and Meina Song. CP-Prompt: Composition-based cross-modal prompting for domain-incremental continual learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28463–28473. IEEE, 2024.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Albert HR Ko, Robert Sabourin, and Alceu Souza Britto Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition*, 41(5):1718–1731, 2008.

Minh Le, Anh Nguyen, Huy Nguyen, Chau Nguyen, Anh T Tran, and Nhat Ho. Revisit visual prompt tuning: The expressiveness of prompt experts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/af01716e08073368a7c8a62be46dba17-Abstract-Conference.html.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Wonseon Lim, Jaesung Lee, and Dae-Won Kim. Critical patch-aware sparse prompting with decoupled training for continual learning on the edge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17989–17998, June 2026.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.

Seoyoung Park, Haemin Lee, and Hankook Lee. Is prompt selection necessary for task-free online continual learning? *arXiv preprint arXiv:2604.04420*, 2026.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JewzobRhay.

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.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6368–6377, 2020.

Ibne Farabi Shihab, Abu Sa-Adat Mohamed Moon-Im Al Ahsan, and Md Najmus Swaqeeb. Opportunity is not realizability: Selection-valid diagnostics for Multi-LLM routing. *arXiv preprint arXiv:2608.08265*, 2026.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2106.10270.

Lukas Thede, Karsten Roth, Olivier J. Hénaff, Matthias Bethge, and Zeynep Akata. Reflecting on the state of rehearsal-free continual learning with pretrained models. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2024.

Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5385–5394. IEEE, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Venkateswara_Deep_Hashing_Network_CVPR_2017_paper.html.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 36:69054–69076, 2023.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, and Jun Zhu. HiDe-PET: Continual learning via hierarchical decomposition of parameter-efficient tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6687–6702, 2025a.

Qiang Wang, Yuhang He, Songlin Dong, Xinyuan Gao, Shaokun Wang, and Yihong Gong. Non-exemplar domain incremental learning via cross-domain concept integration. In *European Conference on Computer Vision*, pp. 144–162. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6534_ECCV_2024_paper.php.

Qiang Wang, Xiang Song, Yuhang He, Jizhou Han, Chenhao Ding, Xinyuan Gao, and Yihong Gong. Boosting domain incremental learning: Selecting the optimal parameters is all you need. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4839–4849. IEEE, 2025b.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-Prompts learning with pre-trained transformers: An Occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022a.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022c.

Ross Wightman. PyTorch Image Models. GitHub repository, 2019. URL https://github.com/huggingface/pytorch-image-models.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14299–14308, 2021.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19148–19158, 2023.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

Supplement roadmap.

- **Setup:** evaluated systems, data, and reconstruction provenance (Appendix A); estimators and removal references (Appendix B).
- **Fixed-bank results:** exhaustive prompt selection (Appendix C), alternate routers including the pre-specified Gaussian router (Appendix D), classifier-scope comparison (Appendix E), and per-domain accounting (Appendix E.1).
- **Classifier exposure:** the matched refit, its controls, and the tied-prompt follow-up (Appendix F).
- **Replication:** further seeds and a reversed domain order (Appendix G); the DualPrompt+PDH banks (Appendix H).
- **Boundaries and context:** limits of the evidence (Appendix I) and relation to prior work (Appendix J).

A Evaluation protocol and implementation provenance

Scope of the completed evidence. The component-selection analysis uses fresh prediction exports from eighteen retained final checkpoints: S-iPrompts, CPrompt, and CP-Prompt, each on Office-Home and DomainNet with training seeds 87, 1013, and 1993. The separate classifier-scope comparison uses three paired training seeds, 0, 1, and 2, per benchmark. These are different experimental families. Three DualPrompt+PDH DomainNet banks (Appendix H) add a fourth evaluated configuration to the four-cell, removal, and exposure analyses. A checkpoint evaluated under several heads, policies, or domains is one trained model, not several independent replications. All analyses here use completed runs; no accuracy from an unfinished training run enters a table or claim.

Shared-label-space domain-incremental evaluation. Office-Home has four domains and 65 classes; DomainNet has six domains and 345 classes (Venkateswara et al., 2017; Peng et al., 2019). The label space is shared across domains. We evaluate the final retained parameter bank, with all domains available to the selector. The true domain is used only in diagnostic interventions and stratification. It is not supplied to the original deployed prediction rule. Office-Home order is Art, Clipart, Product, Real World; DomainNet order is clipart, infograph, painting, quickdraw, real, sketch. These fixed orders are shared within the study and are not an order-robustness experiment.

Office-Home uses a study-defined domain/class-stratified split, with split seed 42: after shuffling each domain/class list, the first $\lfloor 0.8n \rfloor$ images are assigned to training and the remainder to testing. The split is fixed across training seeds. DomainNet retains the official test lists; approximately 10% of each official training partition is reserved for validation. Thus the official test population is retained, but the fitting budget differs from a full-training reproduction. Table 5 gives the scored populations. These are within-study comparisons, not claims to reproduce every original paper's reported benchmark accuracy.

Table 5: **Scored populations and mismatch sets.** Scored test populations and the original routing-mismatch sets. Domain counts are fixed across the evaluated seeds. The mismatch counts depend on the recorded selector. Fixed-bank component and original-mismatch-set policy contrasts retain those assignments; alternate-router analyses explicitly replace them.

Benchmark	Domain	Test samples
Office-Home	Art / Clipart / Product / Real World	507/893/915/898
DomainNet	clipart / infograph / painting	14,604/15,582/21,850
	quickdraw / real / sketch	51,750/52,041/20,916
Family–benchmark	**Training seeds**	**$\lvert\mathcal{M}\rvert$ by seed**
S-iPrompts–OH	87/1013/1993	1,095/1,108/1,054
S-iPrompts–DN	87/1013/1993	34,671/34,715/34,401
CP-Prompt–OH	87/1013/1993	783/828/784
CP-Prompt–DN	87/1013/1993	27,945/28,755/28,816
CPrompt–OH	87/1013/1993	1,083/1,080/1,067
CPrompt–DN	87/1013/1993	29,337/29,383/29,448

Which S-Prompts variant is evaluated. We use the image-only `sip` implementation, denoted S-iPrompts, from S-Prompts (Wang et al., 2022a). It uses an AugReg-pretrained visual ViT-B/16, ten input prompt tokens per domain, separate domain classifiers, and nearest-centroid routing with $K = 5$ domain centroids and the implementation's L_1 rule (Dosovitskiy et al., 2021; Steiner et al., 2022). It is not the CLIP-based language–image variant. For the component analysis, cached features produced by every prompt are passed through every frozen classifier. This constructs the full $N \times T \times T$ grid of class predictions without changing the checkpoint, fitting a new classifier, or rerouting after an intervention.

CPrompt's fixed combined prediction rule. CPrompt is a class-incremental method evaluated here through the shared-label DIL port (Gao et al., 2024). Its domain-indexed prompt bundle contains the port's two task-specific prompt components. Every classifier block remains active. The Office-Home checkpoints are the port's; the DomainNet checkpoints were trained for this study with the port's DomainNet recipe unchanged (six stages in the DomainNet order above, 15 epochs per stage, learning rate 0.01, batch size 16, weight decay 5×10^{-4}, seeds 87, 1013, 1993), with per-stage hashed checkpoints, and exported by the same fresh-prediction procedure with a native-forward parity check on every test image. Given the feature produced by prompt p, our port concatenates the T classifier blocks, whose logits are $z_{j,c}$, and folds repeated class identities:

$$\hat{y}_p(x) = \operatorname*{arg\,max}_{c \in \{1,\ldots,C\}} \log \sum_{j=1}^{T} \exp(z_{j,c}(x; p)) \,. \tag{6}$$

The selector changes the prompt, not a single classifier identity. Consequently CPrompt has one combined-classifier condition in the exhaustive prompt analysis and no native prompt/head 2×2 intervention. Replacing this readout by a selected head or a maximum over heads would evaluate a different predictor.

Fresh prediction reconstruction and provenance. The twelve S-iPrompts and CPrompt exports were reconstructed from the retained checkpoints with explicit sample manifests, stable image identifiers, recorded labels and domains, checkpoint SHA-256 hashes, and the native evaluation preprocessing. The six CP-Prompt evaluations write the same kind of image-indexed export, with a complete sample manifest verified against the training manifest, native forward probes and numerical-health checks, and the training configuration and directory recorded as provenance in the evaluation audit. Each available prompt and the prompt-free path are scored on these same images. S-iPrompts exports the complete prompt-by-head prediction grid and recorded L_1/L_2/cosine assignments; CPrompt exports every prompt under its unchanged combined readout. Native-forward checks and integer repair/harm accounting verify the reconstruction. Primary figures, policy tables, and removal counts are generated from these new exports rather than anonymous historical cache rows.

For DN, historical CPU classifier evaluation and the fresh GPU path differ on eight near-tied entries among 19,088,244 full-grid predictions across three seeds. Replaying the same features and weights under the recorded CPU configuration reproduces those historical predictions; prompted features and routing assignments agree exactly. The primary analysis consistently uses the fresh GPU exports, as do the matched exposure arms. Native domain-correction correctness and fixed-selected/domain-head prompt-oracle correctness agree across the paths. These checks resolve the numerical source chain without claiming that historical row numbers independently identify images. The archived CPU comparison is a numerical reproducibility companion, not a second source for selected primary metrics.

The author-constructed classifier-scope control retains its separate recorded count exports and paired assignment checks (Appendix E); it is not promoted to a fresh image reconstruction by the fresh-export audit. Output and input hashes, source code, and seed-level integer counts accompany the primary export manifest.

Training recipes and preprocessing. The archived configurations use 10 epochs per domain for S-iPrompts on Office-Home and 30 on DomainNet, with batch size 128 and learning rate 0.01. The implementation uses SGD with momentum 0.9 and cosine annealing. Weight decay is 2×10^{-4}, except for DomainNet's first domain, which uses 5×10^{-4}. The DomainNet optimisation

values were adapted from the released language–image configuration while retaining the image-only model; this is not an unchanged reproduction of a released image-only DomainNet recipe. The recorded centroid construction uses normalised prompt-free features and clustering random state zero.

CPrompt's port uses, on both benchmarks, 15 epochs per domain, batch size 16, SGD with learning rate 0.01, momentum 0.9, weight decay 5×10^{-4}, and cosine annealing. Its recorded hyperparameters are $\tau = 1.02$, margin 0.05, $\alpha = 1$, and top-$k = 5$. Two 50-token prompt bundles are inserted, one at the input and one after the first five transformer blocks. CP-Prompt uses, on both benchmarks, a shared prompt of 6 tokens, trained through the stream with a frozen per-domain snapshot kept after each stage and switched together with that domain's prefix, and domain-personalised prefixes of 70 tokens in transformer layers 1–10 (key/value prefix tuning (Li & Liang, 2021)), $K = 5$ nearest-neighbour domain inference, 30 epochs per domain, SGD with learning rate 0.01, batch size 128, and weight decay 2×10^{-4} (5×10^{-4} for the first domain); its classifier is the set of class prototypes generated from learnable text context. Both S-iPrompts and CPrompt train with random resized 224×224 crops and horizontal flips, and evaluate with resize-to-256 followed by a centre crop of 224. Their normalisations differ: S-iPrompts uses ImageNet channel statistics, whereas the CPrompt port parameterises its normalisation with mean 0.0 and standard deviation 1.0, which acts as an identity and leaves pixel tensors in $[0, 1]$ rather than applying dataset z-scoring. These implementation choices are preserved rather than treated as controlled across methods.

Artifact access. The source package contains the frozen numerical exports, training-configuration provenance, and analysis/plotting code. It supports regeneration of the reported summaries and figures; it does not contain the large checkpoint and image-level prediction archives needed to rerun every reconstruction. The figure and table generation code, together with the per-seed summaries it consumes, is available at `https://anonymous.4open.science/r/prompt-head-decomposition-F1C1/README.md`.

B PAIRED ESTIMATORS, DENOMINATORS, AND REMOVAL REFERENCES

Accuracy changes and event counts. For a scored set S of size $|S|$, let $\hat{y}_i$ and $\hat{y}'_i$ denote the predictions before and after a specified intervention. Define repair and harm counts

$$R_S = \sum_{i \in S} \mathbf{1}[\hat{y}_i \neq y_i,\ \hat{y}'_i = y_i], \qquad H_S = \sum_{i \in S} \mathbf{1}[\hat{y}_i = y_i,\ \hat{y}'_i \neq y_i]. \tag{7}$$

The accuracy change in percentage points is exactly $100(R_S - H_S)/|S|$. Changed predictions number at least $R_S + H_S$; the remainder comprises changed but still incorrect predictions. No class probability or calibration assumption is needed for these identities.

Let $\mathcal{M} = \{i : s_i \neq d_i\}$ be the mismatch set of the recorded original selector. Domain correction with a fixed retained head has identical predictions outside $\mathcal{M}$. Its conditional and full-population changes, with N the full test size, obey

$$\Delta_{\text{full}} = \frac{|\mathcal{M}|}{N} \Delta_{\mathcal{M}} = 100 \frac{R_{\mathcal{M}} - H_{\mathcal{M}}}{N}. \tag{8}$$

This gives the familiar routing-gain decomposition by mismatch mass and conditional net repair. If $\mathcal{M}$ is empty, the full change is zero while the conditional rate is undefined. Prompt removal and prompt-oracle selection can also change predictions outside $\mathcal{M}$; their full-population changes cannot be recovered from the mismatch term alone. Domain-level rates use N_d or $|\mathcal{M}_d|$, as indicated. Counts, rather than unweighted averages of domain rates, reconstruct pooled micro-accuracies.

Prediction changes on the original mismatch set. For prompt correction under the selected head, the prediction-change rate is

$$100 \, \frac{\#\{i \in \mathcal{M} : \hat{y}_{ds,i} \neq \hat{y}_{ss,i}\}}{|\mathcal{M}|}.$$

In S-iPrompts DN, the changed-prediction counts are 9,490, 9,702, and 9,920 for seeds 87, 1013, and 1993, with mismatch counts 34,671, 34,715, and 34,401. Their paired rates average

$28.052 \pm 0.738\%$. Repairs and harms account for 3.584% and 5.101% of $\mathcal{M}$ on average; the remaining 19.367% are changed but still wrong predictions. These entries are recorded in the released `data/displayed_evidence.json`, for S-iPrompts/DN/selected head and the domain-correction policy on $\mathcal{M}$. The displayed net gain counts only repairs minus harms, not all changed predictions.

Paired aggregation and uncertainty. We form every contrast separately within each training seed and report its arithmetic mean and sample standard deviation, with denominator $3 - 1$. For example, the interaction's SD is computed from three within-seed interaction values, not from independently propagated SDs of its constituent accuracies. Equation 8 is applied before seed averaging; a product of rounded mean mismatch mass and rounded mean conditional gain need not reproduce a displayed mean. Domain subdivisions and the two alternate routing rules reuse the same trained models. They increase diagnostic resolution, but do not increase the number of independent training replications. The reported SD is descriptive training-seed variability, not a confidence interval or a formal population-level significance test.

Table 2 in the main text reports the four-cell contrasts; this appendix specifies their references and paired estimators.

The retained head and reference prompt are separate choices. The two prompt corrections are $A_{ds} - A_{ss}$ and $A_{dd} - A_{sd}$; their difference is the interaction I_{PH} of Equation 2. A nonzero value means that the measured replacement effect of one component depends on the other component's selected state. It does not assign a unique percentage of model knowledge to either component or identify the cause of the interaction during training.

The body reports removal under the deployed head, $A_{\varnothing s} - A_{ss}$: the selected prompt is removed while the selected head is retained (-23.762 ± 1.301 pp on DN and -4.720 ± 1.716 pp on OH for S-iPrompts). The archived removal export and Figure 5 use $A_{\varnothing d} - A_{dd}$: the domain-matched prompt is removed while the domain head is retained. In that export and figure, CP-Prompt removal uses the same contrast: all visual prompts, the common prompt and the personalised prefixes, are removed while the domain head, its text-derived prototypes, is retained, and the reference is the domain prompt. Its endpoint and baseline differ from domain correction, even when the retained head is the same. Removal measures reliance of the trained predictor under this intervention; it does not test whether a model trained without prompts can reach the same accuracy.

Removal is recomputed per image from the fresh prompt-free and domain-prompt predictions, with the same head and image manifest. The released removal export records N, both correct counts, repair, harm, and the signed net change for every seed, dataset, and reference. The displayed 4.63 ± 1.71 and 24.66 ± 1.32 pp removal costs therefore use integer outcomes rather than rounded historical domain accuracies. Domain-weighted micro totals and equal-domain averages remain distinct.

CPrompt's recorded current accuracy is $89.625 \pm 0.177\%$, prompt-free accuracy $88.847 \pm 0.118\%$, and domain-prompt accuracy $90.072 \pm 0.325\%$. Thus removing prompts relative to learned selection and removing them relative to the domain prompt have different baselines: approximately -0.778 and -1.224 pp, respectively. The combined readout is unchanged in both cases. On DomainNet the same three quantities are $70.415 \pm 0.061\%$, $51.909 \pm 0.267\%$, and $70.768 \pm 0.054\%$, i.e. -18.506 and -18.859 pp. The fresh removal export records both selected-prompt and domain-prompt references explicitly.

C Exhaustive prompt selection and the mismatch partition

Candidate predictions and the hard-selection oracle. The label-informed sample-wise oracle

$$O_h(S) = \frac{1}{|S|} \sum_{(x,y,d) \in S} \max_{p \in \{1,\dots,T\}} \mathbf{1}[f(x; p, h(x)) = y] - A_h(s; S) \tag{9}$$

includes the current prompt, so $O_h \geq \max(0, C_h)$ for a finite bank and fixed head. In S-iPrompts on DN under the domain head it is 6.654 ± 0.296 pp, where domain correction gains 0.472 ± 0.076 pp (Table 6 and Figure 5 in Appendix C).

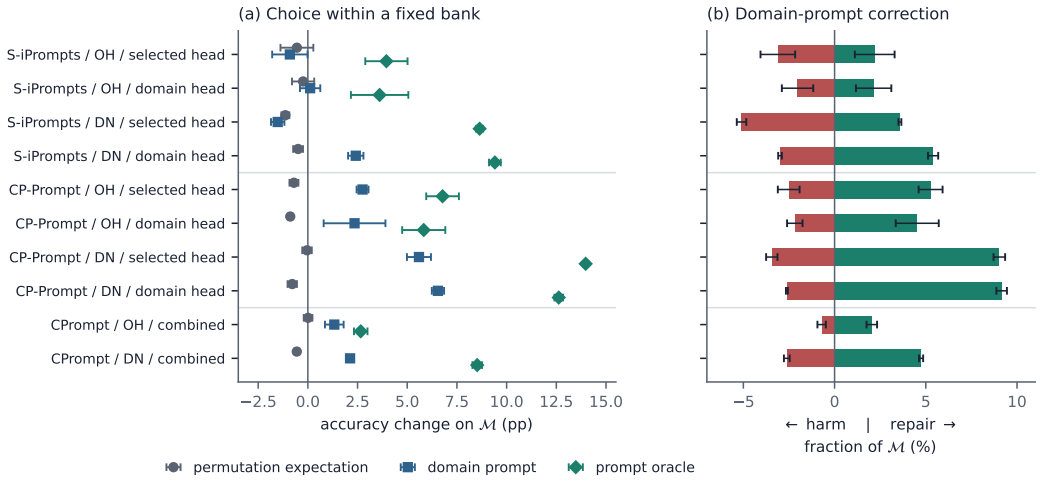

Figure 4: **Disambiguating a small correction gain on the original mismatch set.** All comparisons use $\mathcal{M} = \{s \neq d\}$ with the head, or CPrompt's combined readout, held fixed per example. (a) Accuracy changes relative to the original selected prompt. The permutation point is an exact within-domain, frequency-preserving reassignment expectation; the oracle uses test labels. (b) Domain correction repairs some predictions and harms others; harm is drawn on the negative axis. Bars and whiskers show three-seed means and sample SDs. The ten rows are conditions of eighteen checkpoints, not ten independent methods.

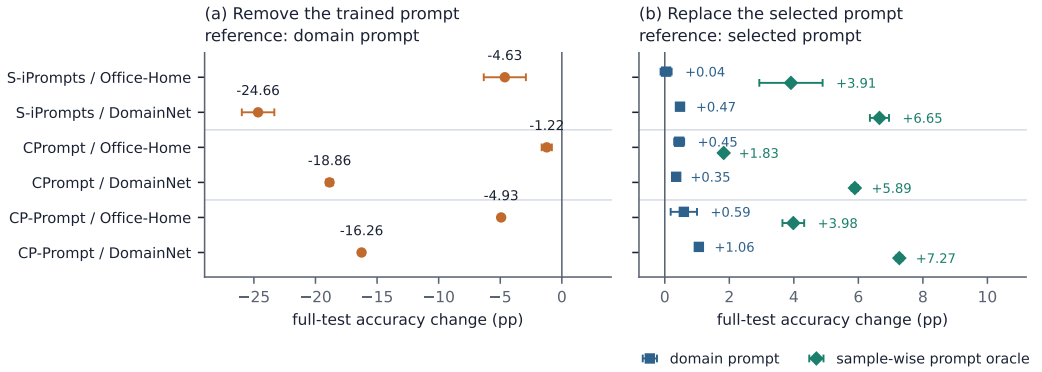

Figure 5: **Prompt dependence, domain correction, and selection opportunity answer different questions.** Final-stage, full-test accuracy changes (three training seeds; bars show sample SD). S-iPrompts and CP-Prompt use a fixed domain head; CPrompt uses its fixed combined readout. (a) Removing the prompt is measured relative to the domain prompt. (b) Domain correction and the label-informed sample-wise prompt oracle are both measured relative to the selected prompt. The two references are intentionally distinguished: removal and correction are not additive component attributions. The oracle shares correction's reference and population; it describes available alternatives without establishing domain specialisation or a deployable gain.

For each input, hold its readout fixed and let q_{ip} be the class prediction using prompt p. S-iPrompts uses either h_{s_i} or h_{d_i} throughout a given comparison; CPrompt uses the combined readout above. The baseline is q_{i,s_i}. Domain correction uses q_{i,d_i}. The retrospective prompt oracle is correct if and only if at least one of the T candidate prompts predicts y_i:

$$c_i^\star = \max_{p \in \{1,\ldots,T\}} \mathbf{1}[q_{ip} = y_i]. \tag{10}$$

Because the candidate set includes the current prompt, this oracle never harms baseline correctness. It is defined here as an oracle correctness event, not as a unique predicted label on all-wrong inputs.

Table 6: **Domain correction versus available prompt selection on all test examples, the eighteen original checkpoints.** Current accuracy is in percent; changes are pp relative to the current selected prompt with the indicated fixed readout. The oracle chooses among all T existing prompts using labels. OH: $N = 3{,}213$; DN: $N = 176{,}743$. Mean ± sample SD over three training seeds.

Family	Data	Readout	Current acc.	Domain correction	Prompt oracle
S-iPrompts	OH	h_s	81.71 ± 1.33	-0.301 ± 0.289	$+4.025 \pm 0.856$
S-iPrompts	OH	h_d	84.11 ± 1.33	$+0.041 \pm 0.171$	$+3.911 \pm 0.981$
S-iPrompts	DN	h_s	61.83 ± 1.12	-0.297 ± 0.064	$+6.503 \pm 0.383$
S-iPrompts	DN	h_d	65.49 ± 1.16	$+0.472 \pm 0.076$	$+6.654 \pm 0.296$
CP-Prompt	OH	h_s	88.01 ± 0.35	$+0.685 \pm 0.093$	$+4.212 \pm 0.234$
CP-Prompt	OH	h_d	88.65 ± 0.43	$+0.591 \pm 0.408$	$+3.984 \pm 0.337$
CP-Prompt	DN	h_s	73.11 ± 0.12	$+0.902 \pm 0.106$	$+7.488 \pm 0.114$
CP-Prompt	DN	h_d	74.24 ± 0.16	$+1.056 \pm 0.067$	$+7.270 \pm 0.092$
CPrompt	OH	combined	89.63 ± 0.18	$+0.446 \pm 0.154$	$+1.826 \pm 0.048$
CPrompt	DN	combined	70.41 ± 0.06	$+0.353 \pm 0.024$	$+5.891 \pm 0.099$

Prediction-change rates for the oracle are therefore left undefined. The oracle bounds hard selection among these fixed candidates with this fixed readout. It does not bound arbitrary aggregation, a different bank, or a retrained predictor, and it is not an estimate of deployable router accuracy.

Frequency-preserving permutation control. Within true domain d, consider only its original mismatched inputs $\mathcal{M}_d$. Let n_{dp} count how many of them the recorded router assigns to prompt p, and set $w_{dp} = n_{dp}/|\mathcal{M}_d|$. The reported control is the exact expected outcome of uniformly permuting those prompt assignments across $\mathcal{M}_d$. For example, its expected accuracy is

$$\frac{1}{|\mathcal{M}|} \sum_{d} \sum_{i \in \mathcal{M}_d} \sum_{p} w_{dp} \mathbf{1}[q_{ip} = y_i]. \tag{11}$$

Expected repairs and harms are computed by replacing the final correctness indicator with the corresponding paired event indicator. Expected counts can be fractional. This calculation retains prompt-selection frequencies and each sample's fixed retained readout while breaking the association between the input and the assigned prompt. It includes permutations that return an input to its original prompt. It is not a forced change to a different prompt, a logit ensemble, or a Monte Carlo estimate. The displayed SD comes from training seeds, not permutation sampling. The control is defined only on $\mathcal{M}$; no full-test permutation accuracy is imputed.

Under selected heads, permutation additionally breaks the original prompt–head identity pairing. Its loss does not isolate input-specific prompt selection from that pairing effect. Under domain heads, the head is constant within each true-domain permutation stratum, making this condition the more direct test of the input–prompt association.

Table 7: **All mismatch-set policies.** Complete pooled mismatch-set comparisons. Repair and harm are percentages of the original mismatch set under domain correction. The last three columns are accuracy changes on $\mathcal{M}$, in pp relative to the selected prompt under the stated fixed readout. Permutation values are exact expectations; all entries report mean and sample SD across three training seeds.

Family	Data	Readout	Repair	Harm	Domain	Permutation	Oracle
S-iPrompts	OH	h_s	2.20 ± 1.09	3.11 ± 0.95	-0.91 ± 0.89	-0.55 ± 0.83	$+3.95 \pm 1.06$
S-iPrompts	OH	h_d	2.14 ± 0.97	2.03 ± 0.86	$+0.11 \pm 0.51$	-0.24 ± 0.56	$+3.61 \pm 1.45$
S-iPrompts	DN	h_s	3.58 ± 0.08	5.10 ± 0.26	-1.52 ± 0.34	-1.13 ± 0.20	$+8.64 \pm 0.18$
S-iPrompts	DN	h_d	5.40 ± 0.28	2.99 ± 0.10	$+2.41 \pm 0.39$	-0.49 ± 0.25	$+9.41 \pm 0.30$
CP-Prompt	OH	h_s	5.26 ± 0.66	2.51 ± 0.60	$+2.75 \pm 0.30$	-0.70 ± 0.22	$+6.78 \pm 0.82$
CP-Prompt	OH	h_d	4.53 ± 1.18	2.18 ± 0.42	$+2.35 \pm 1.55$	-0.89 ± 0.11	$+5.83 \pm 1.09$
CP-Prompt	DN	h_s	9.03 ± 0.32	3.44 ± 0.31	$+5.59 \pm 0.60$	-0.04 ± 0.24	$+13.98 \pm 0.14$
CP-Prompt	DN	h_d	9.16 ± 0.29	2.61 ± 0.05	$+6.55 \pm 0.31$	-0.79 ± 0.24	$+12.62 \pm 0.22$
CPrompt	OH	combined	2.04 ± 0.29	0.71 ± 0.23	$+1.33 \pm 0.47$	$+0.01 \pm 0.22$	$+2.66 \pm 0.34$
CPrompt	DN	combined	4.74 ± 0.11	2.62 ± 0.16	$+2.12 \pm 0.15$	-0.56 ± 0.12	$+8.52 \pm 0.24$

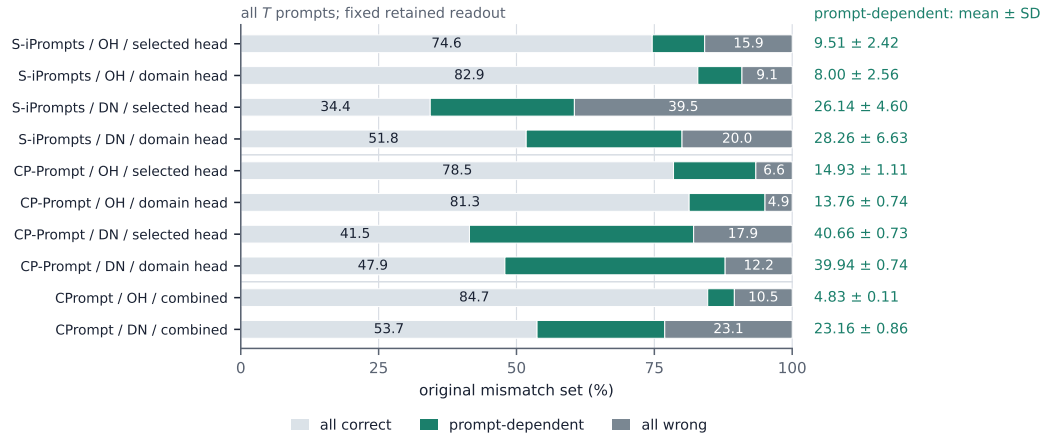

Figure 6: **Correctness varies across prompts for a subset of inputs.** Fractions of the original mismatch set for the eighteen original checkpoints, scoring all T prompts under a fixed readout. Bars show three-seed means; the adjacent column shows the prompt-dependent fraction's mean $\pm$ sample SD. The categories partition each seed's inputs. A prompt-dependent input, correct under some prompts and not others, is required for selection improvement but may already be correct under the selected prompt; the fraction is not the oracle gain. All-wrong prompts may disagree among incorrect labels. All category SDs appear in Table 8.

Table 8: **Where can prompt choice change correctness?** Percentages of the original mismatch set, mean $\pm$ sample SD over three seeds, with all T prompts scored under a fixed readout. The three categories partition $\mathcal{M}$ within each seed. Prompt-dependent inputs, correct under some prompts and not others, are the only ones selection can improve, but the category does not itself mean that the current prompt is wrong.

Family	Data	Readout	All correct	Prompt-dependent	All wrong
S-iPrompts	OH	h_s	74.62 ± 2.76	9.51 ± 2.42	15.87 ± 0.50
S-iPrompts	OH	h_d	82.89 ± 2.70	8.00 ± 2.56	9.12 ± 0.16
S-iPrompts	DN	h_s	34.36 ± 4.66	26.14 ± 4.60	39.50 ± 0.27
S-iPrompts	DN	h_d	51.76 ± 6.54	28.26 ± 6.63	19.99 ± 0.09
CP-Prompt	OH	h_s	78.48 ± 1.04	14.93 ± 1.11	6.60 ± 0.07
CP-Prompt	OH	h_d	81.31 ± 0.93	13.76 ± 0.74	4.92 ± 0.41
CP-Prompt	DN	h_s	41.47 ± 0.94	40.66 ± 0.73	17.87 ± 0.43
CP-Prompt	DN	h_d	47.90 ± 0.84	39.94 ± 0.74	12.16 ± 0.16
CPrompt	OH	combined	84.68 ± 0.69	4.83 ± 0.11	10.49 ± 0.65
CPrompt	DN	combined	53.73 ± 0.84	23.16 ± 0.86	23.11 ± 0.08

Relation to the correctness partition. Figure 6 reports the proportions correct under every prompt, under some prompts, or under none. These categories sum to one within each seed. The oracle's repairs are a subset of the prompt-dependent category: inputs for which the current prompt is incorrect but at least one alternative is correct. This relation is checked in the figure/table generation code for every reported readout.

Opportunity outside domain mismatches. Table 9 reports both conditional oracle gains and their contributions to the entire test population. Let $G_{\mathcal{M}}$ and $G_{\overline{\mathcal{M}}}$ be oracle gains on their own conditional denominators. Then

$$G_{\text{all}} = \frac{|\mathcal{M}|}{N} G_{\mathcal{M}} + \frac{N - |\mathcal{M}|}{N} G_{\overline{\mathcal{M}}}. \tag{12}$$

For S-iPrompts DomainNet under domain heads, the terms contribute 1.842 and 4.812 pp to the 6.654 pp full-test opportunity. Thus most of the full-test oracle gain lies outside the domain-mismatch set, where domain correction is identically zero. The selected-head and domain-head

rows agree outside $\mathcal{M}$ because $s_i = d_i$ there. This accounting distinguishes the domain-correction diagnostic from the broader fixed-bank selection problem.

Table 9: **Where the oracle opportunity lies.** Prompt-oracle opportunity partition. Conditional columns use the size of the indicated subset; contribution columns divide the same net counts by full test size N. The two contributions sum to the full-test oracle gain within each seed. Their SDs must not be added.

			Conditional gain (pp)		Contribution to full gain (pp)	
Family	Data	Readout	$\mathcal{M}$	$\overline{\mathcal{M}}$	$\mathcal{M}$	$\overline{\mathcal{M}}$
S-iPrompts	OH	h_s	$+3.95 \pm 1.06$	$+4.06 \pm 0.78$	$+1.338 \pm 0.382$	$+2.687 \pm 0.514$
S-iPrompts	OH	h_d	$+3.61 \pm 1.45$	$+4.06 \pm 0.78$	$+1.224 \pm 0.512$	$+2.687 \pm 0.514$
S-iPrompts	DN	h_s	$+8.64 \pm 0.18$	$+5.98 \pm 0.44$	$+1.691 \pm 0.037$	$+4.812 \pm 0.356$
S-iPrompts	DN	h_d	$+9.41 \pm 0.30$	$+5.98 \pm 0.44$	$+1.842 \pm 0.060$	$+4.812 \pm 0.356$
CP-Prompt	OH	h_s	$+6.78 \pm 0.82$	$+3.37 \pm 0.22$	$+1.681 \pm 0.165$	$+2.531 \pm 0.147$
CP-Prompt	OH	h_d	$+5.83 \pm 1.09$	$+3.37 \pm 0.22$	$+1.452 \pm 0.302$	$+2.531 \pm 0.147$
CP-Prompt	DN	h_s	$+13.98 \pm 0.14$	$+6.24 \pm 0.11$	$+2.254 \pm 0.045$	$+5.233 \pm 0.091$
CP-Prompt	DN	h_d	$+12.62 \pm 0.22$	$+6.24 \pm 0.11$	$+2.036 \pm 0.068$	$+5.233 \pm 0.091$
CPrompt	OH	combined	$+2.66 \pm 0.34$	$+1.40 \pm 0.24$	$+0.892 \pm 0.109$	$+0.934 \pm 0.156$
CPrompt	DN	combined	$+8.52 \pm 0.24$	$+5.37 \pm 0.07$	$+1.417 \pm 0.038$	$+4.475 \pm 0.061$

The full-test domain-head oracle gap above the domain-matched prompt is 6.182 ± 0.370 pp on DomainNet. On the mismatch set alone, that residual is 6.999 ± 0.091 pp, contributing 1.370 ± 0.019 pp to the full population. These are paired oracle-minus-domain contrasts, not the oracle gains relative to the selected-prompt baseline in Table 7. The distinction prevents the same number from being assigned to two different reference predictors.

D Component effects under alternate test-time routers

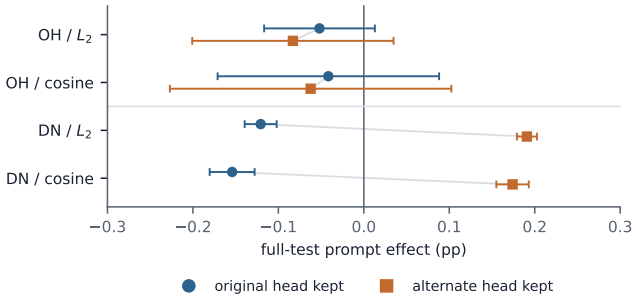

Figure 7: **Retained-head dependence under router changes requiring no test-domain labels.** Full-test accuracy change in pp when only the prompt assignment changes from native L_1 to L_2 or cosine, holding the native-router selected head (circles) or alternate-router selected head (squares) fixed; each point is measured against that same head's accuracy under the native prompt. Mean $\pm$ sample SD over three S-iPrompts seeds. The DomainNet sign reversal occurs in every seed; the Office-Home means stay at or below zero under either head.

To test whether retained-head dependence also appears under ordinary route changes, we retain the archived native L_1, L_2, and cosine assignments for each S-iPrompts and CP-Prompt bank. No candidate is fitted or selected using the resulting test accuracies in this reanalysis. The three rules route $80.426 \pm 0.096\%$, $77.967 \pm 0.076\%$, and $77.894 \pm 0.096\%$ of the DomainNet test set to its true domain, so the alternates cost 2.459 ± 0.027 and 2.532 ± 0.029 pp of routing accuracy, in every seed. Let s_i be the native assignment and r_i an alternate assignment. The four cells use (p_{s_i}, h_{s_i}), (p_{r_i}, h_{s_i}), (p_{s_i}, h_{r_i}), and (p_{r_i}, h_{r_i}). All cells agree outside $\mathcal{D} = \{i : s_i \neq r_i\}$. The published contrasts below use full-test N, while the data also retain $\mathcal{D}$-conditional and domain-level rows.

Table 10 reports every alternate-router component contrast, and Figure 7 shows the prompt-only changes. For S-iPrompts, the changed-assignment counts for seeds 87/1013/1993 are 407/399/405 for Office-Home $L_1 \to L_2$, 459/440/428 for Office-Home $L_1 \to$ cos, 15,187/15,136/15,124 for

Table 10: **Alternate-router cells.** Complete alternate-router component contrasts for S-iPrompts and CP-Prompt, in full-test pp. Prompt only is $A_{rs} - A_{ss}$ under h_s and $A_{rr} - A_{sr}$ under h_r; head only is $A_{sr} - A_{ss}$; I_{PH} and $\Delta J = A_{rr} - A_{ss}$ are as defined in the body. Mean $\pm$ sample SD over three training seeds; the identities are checked within each seed.

Family	Data	Route change	Prompt (h_s)	Prompt (h_r)	Head only	I_{PH}	ΔJ
S-iPrompts	OH	$L_1 \to L_2$	-0.052 ± 0.065	-0.083 ± 0.118	-0.539 ± 0.072	-0.031 ± 0.062	-0.622 ± 0.189
S-iPrompts	OH	$L_1 \to \cos$	-0.041 ± 0.130	-0.062 ± 0.165	-0.664 ± 0.078	-0.021 ± 0.072	-0.726 ± 0.242
S-iPrompts	DN	$L_1 \to L_2$	-0.121 ± 0.019	$+0.191 \pm 0.012$	-0.812 ± 0.019	$+0.312 \pm 0.030$	-0.621 ± 0.016
S-iPrompts	DN	$L_1 \to \cos$	-0.154 ± 0.026	$+0.174 \pm 0.019$	-0.875 ± 0.042	$+0.328 \pm 0.038$	-0.701 ± 0.026
CP-Prompt	OH	$L_1 \to L_2$	-0.052 ± 0.160	$+0.000 \pm 0.136$	-0.073 ± 0.048	$+0.052 \pm 0.095$	-0.073 ± 0.180
CP-Prompt	OH	$L_1 \to \cos$	-0.062 ± 0.082	$+0.062 \pm 0.112$	-0.104 ± 0.065	$+0.124 \pm 0.143$	-0.041 ± 0.130
CP-Prompt	DN	$L_1 \to L_2$	-0.105 ± 0.017	-0.100 ± 0.031	-0.127 ± 0.051	$+0.006 \pm 0.045$	-0.227 ± 0.020
CP-Prompt	DN	$L_1 \to \cos$	-0.143 ± 0.011	-0.136 ± 0.040	-0.111 ± 0.037	$+0.006 \pm 0.046$	-0.248 ± 0.006

DomainNet $L_1 \to L_2$, and $16{,}588/16{,}348/16{,}219$ for DomainNet $L_1 \to \cos$. In the S-iPrompts DomainNet banks the prompt-only change is negative with the original head and positive with the alternate head in all three seeds for both route changes, whereas in the CP-Prompt DomainNet banks it stays negative under either head. Specifically, the S-iPrompts alternate-head prompt-only change is $0.1828/0.1856/0.2043$ pp for L_2 and $0.1522/0.1839/0.1861$ pp for cosine.

This is evidence that the observed component dependence is not restricted to an intervention indexed by the ground-truth domain. Both alternate intact-branch policies, however, reduce accuracy relative to native L_1, and the two changes reuse the same features, classifiers, and training seeds rather than replicating across methods. Office-Home does not reproduce the DomainNet sign reversal consistently; we report both benchmarks and retain that boundary.

A protocol-locked better router. Before scoring any candidate on test data, we fixed a router expected to improve domain identification, three descriptive comparators, the contrasts, and the decision rules. The primary router fits a Gaussian to each domain's own prompt-free training features, with a Ledoit–Wolf covariance and equal priors, and routes each test image by maximum likelihood; it uses no test-domain labels. Because the prompt-free encoder is frozen, each domain's statistics come from that domain's training data alone and no earlier samples are stored. The comparators are the same fits without the log-determinant (Mahalanobis distance), the native $K{=}5$ keys refitted on the same cached training features, and the branch whose own head is most confident. Table 11 reports every candidate. These values use the cached features of Appendix G; recomputing the native, L_2, cosine, and domain-oracle cells from them reproduces the published values to within 0.001 pp and every changed-assignment count exactly.

On DomainNet the Gaussian router raises routing accuracy from $80.426 \pm 0.096\%$ to 83.296% and routed accuracy by 1.277 ± 0.045 pp. Its prompt-only change is -0.225 ± 0.029 pp under the native heads and $+0.371 \pm 0.074$ pp under its own. Exact McNemar tests within each fitted bank give $p < 10^{-145}$ for both accuracy gains, and $p \le 1.7 \times 10^{-13}$ and $p < 1.5 \times 10^{-38}$ for the two prompt-only changes, in every primary seed. The locked criterion (higher routing and routed accuracy, then a negative native-head and a positive own-head prompt-only change, each with $p < 0.05$ in all three primary seeds) is therefore met, and the two added canonical-order seeds also meet it (all $p < 10^{-16}$). As with the oracle swap, the reassignments that correct the domain lower accuracy under the native head in every seed, by 1.30% of those images on average; reassignments that break a correct route lower it by 2.94%.

The comparators locate the gain. Without the log-determinant, the same statistics route only 47.3% of images correctly; this router keeps the native-head loss but not a consistent own-head gain. The refitted keys route 79.9%, so the improvement is attributable mainly to the likelihood rule rather than to the cached feature pipeline. Branch confidence is a poor domain selector (41.1%). The Gaussian router depends only on frozen features, so its assignments are identical across training seeds, and seed variation in its cells reflects the prompts and heads. As a probe rather than a proposed router, it stores 1.78M numbers, against 23k for the native keys and 1.59M for the six heads. On Office-Home it raises routing accuracy by 13.4 pp and routed accuracy by 1.43 ± 0.33 pp, but its native-head prompt-only change varies in sign across seeds (-0.249 ± 0.297 pp), in line with the Office-Home boundary above.

Table 11: **Pre-specified router comparison, S-iPrompts.** Full-test pp except routing accuracy (RA, %); mean ± sample SD over training seeds 87/1013/1993. Prompt only is $A_{rs} - A_{ss}$ under h_s and $A_{rr} - A_{sr}$ under h_r; head only is $A_{sr} - A_{ss}$; I_{PH} and $\Delta J = A_{rr} - A_{ss}$ are as defined in the body, relative to native L_1. Gaussian is the primary router; Mahalanobis uses the same fits without the log-determinant, Refit L_1 the native rule with refitted keys, and Confidence the most confident branch. Gaussian, Mahalanobis, and Refit L_1 use frozen features only, so their RA does not vary across seeds. The L_2 and cosine rows repeat Table 10, and the domain-oracle rows repeat Table 2.

Data	Router	RA	Prompt (h_s)	Prompt (h_r)	Head only	I_{PH}	ΔJ
DN	Native	80.426 ± 0.096					
DN	Gaussian	83.296 ± 0.000	-0.225 ± 0.029	$+0.371 \pm 0.074$	$+0.907 \pm 0.044$	$+0.596 \pm 0.082$	$+1.277 \pm 0.045$
DN	Mahalanobis	47.275 ± 0.000	-10.075 ± 1.094	$+0.092 \pm 1.133$	-14.161 ± 0.139	$+10.167 \pm 2.156$	-14.069 ± 1.208
DN	Refit L_1	79.949 ± 0.000	-0.117 ± 0.018	$+0.165 \pm 0.013$	-0.269 ± 0.041	$+0.283 \pm 0.026$	-0.104 ± 0.053
DN	Confidence	41.068 ± 3.367	-6.113 ± 1.245	$+0.382 \pm 0.282$	-6.015 ± 1.282	$+6.495 \pm 1.526$	-5.633 ± 1.001
DN	L_2	77.967 ± 0.076	-0.121 ± 0.019	$+0.191 \pm 0.012$	-0.812 ± 0.019	$+0.312 \pm 0.030$	-0.621 ± 0.016
DN	Cosine	77.894 ± 0.096	-0.154 ± 0.026	$+0.174 \pm 0.019$	-0.875 ± 0.042	$+0.328 \pm 0.038$	-0.701 ± 0.026
DN	Domain oracle	100.000 ± 0.000	-0.297 ± 0.064	$+0.472 \pm 0.076$	$+3.669 \pm 0.039$	$+0.769 \pm 0.117$	$+4.141 \pm 0.038$
OH	Native	66.210 ± 0.877					
OH	Gaussian	79.645 ± 0.000	-0.249 ± 0.297	$+0.145 \pm 0.078$	$+1.286 \pm 0.250$	$+0.394 \pm 0.265$	$+1.432 \pm 0.328$
OH	Mahalanobis	63.399 ± 0.000	-0.550 ± 0.688	-0.249 ± 0.452	-0.944 ± 0.190	$+0.301 \pm 0.253$	-1.193 ± 0.588
OH	Refit L_1	65.359 ± 0.000	-0.166 ± 0.212	-0.010 ± 0.118	$+0.207 \pm 0.238$	$+0.156 \pm 0.329$	$+0.197 \pm 0.229$
OH	Confidence	48.221 ± 2.405	-0.446 ± 0.522	$+0.591 \pm 0.863$	-0.591 ± 0.189	$+1.037 \pm 0.341$	$+0.000 \pm 0.674$
OH	L_2	62.745 ± 1.121	-0.052 ± 0.065	-0.083 ± 0.118	-0.539 ± 0.072	-0.031 ± 0.062	-0.622 ± 0.189
OH	Cosine	62.340 ± 1.303	-0.041 ± 0.130	-0.062 ± 0.165	-0.664 ± 0.078	-0.021 ± 0.072	-0.726 ± 0.242
OH	Domain oracle	100.000 ± 0.000	-0.301 ± 0.289	$+0.041 \pm 0.171$	$+2.397 \pm 0.112$	$+0.342 \pm 0.124$	$+2.438 \pm 0.279$

The same router under a combined readout. CPrompt's readout combines all blocks, so a routing change alters only the prompt; there, a prompt change is the deployed change. We applied the same protocol-locked router to the prompt-free query features that CPrompt's native key router reads, under an analogous gate that reproduced the exported native routes on every test image and the published accuracy and domain correction. On DomainNet the Gaussian router raises routing accuracy from 83.37% to 85.78% and accuracy under the unchanged readout by 0.189 ± 0.025 pp, positive in every seed (exact McNemar $p \leq 2.2 \times 10^{-10}$), and, as with the domain correction, its domain-correcting reassignments raise accuracy on those images in every seed (3.1% on average); Mahalanobis again collapses (59.7% routing), and both routers' assignments are identical across seeds (Table 12). As a probe, it stores 1.78M numbers, against 1.59M for CPrompt's 2,070 native keys. The same kind of router thus raises accuracy through its prompt choices under CPrompt's combined readout and lowers it under S-iPrompts' retained heads; because the two families also differ in backbone, preprocessing, and training, this contrast indicates rather than isolates the role of the readout.

Table 12: **Pre-specified router comparison, CPrompt combined readout (DomainNet).** Routing accuracy (RA, %) and full-test accuracy change in pp when only the prompt assignment changes under the fixed log-sum-exp readout, relative to the native key router; mean ± sample SD over training seeds 87/1013/1993. Gaussian and Mahalanobis use frozen features only, so their RA does not vary across seeds. The domain-oracle row is the domain correction reported in Section 4.1.

Router	RA	ΔRA	Prompt change
Native key router	83.372 ± 0.032		
Gaussian	85.775 ± 0.000	$+2.404 \pm 0.032$	$+0.189 \pm 0.025$
Mahalanobis	59.747 ± 0.000	-23.625 ± 0.032	-2.998 ± 0.073
Domain oracle	100.000 ± 0.000	$+16.628 \pm 0.032$	$+0.353 \pm 0.024$

E CLASSIFIER-SCOPE COMPARISON AND DOMAIN HETEROGENEITY

Correction value under common assignments. The fixed-checkpoint results ask what replacing a component does *after training*. A complementary control asks whether the same assignments have the same predictive cost in different trained banks. We compare our shared-classifier and domain-classifier substrate variants using a common recorded router. Each reference router supplies the assignments for *both* banks, so their domain-selection accuracy and mismatch set are exactly the

same within a paired comparison. We repeat the comparison using the shared-bank router and the domain-bank router as references.

Table 13 reports the change in joint domain-correction effects, domain-classifier bank minus shared-classifier bank. Under the shared-bank router, net correction value increases by 2.531 ± 0.376 pp on OH and 2.682 ± 0.211 pp on DN. Repairs increase more than harms; neither quantity can be recovered from domain-selection accuracy alone. The contrast remains positive for every paired seed and for either reference router; the DomainNet contrast is interpreted alongside the retention check below, while the Office-Home contrast is limited by that control. The two references reuse the same banks and are a robustness check, not independent training replications.

Table 13: **Classifier-scope variants differ in domain-correction value at fixed assignments.** Domain-classifier minus shared-classifier bank, full-test pp, paired mean $\pm$ sample SD. Repair and harm are full-test event fractions multiplied by 100; net equals repair minus harm within each seed. A positive harm contrast means that correction also damages more previously correct predictions.

Data	Common router	Repair change	Harm change	Net change
OH	shared-bank	$+3.299 \pm 0.173$	$+0.768 \pm 0.299$	$+2.531 \pm 0.376$
OH	domain-bank	$+3.185 \pm 0.188$	$+0.778 \pm 0.165$	$+2.407 \pm 0.350$
DN	shared-bank	$+3.238 \pm 0.237$	$+0.556 \pm 0.404$	$+2.682 \pm 0.211$
DN	domain-bank	$+3.261 \pm 0.253$	$+0.556 \pm 0.394$	$+2.705 \pm 0.233$

This control establishes a difference in correction costs under identical routing assignments. It does not make the banks equally accurate or identify a single training mechanism. Their native accuracies differ, and changing classifier scope changes the learned system. Domain-level effects are also heterogeneous: for the shared reference on DN, the infograph contrast is -1.671 ± 2.817 pp even though the pooled contrast is positive. We therefore do not equate this controlled training comparison with the S-iPrompts substitution experiment, nor conclude that domain-specific heads are necessary or sufficient for large selection value.

Training contrast and common routing. The author-constructed substrate uses a frozen ViT-B/16 backbone and ten domain-specific input prompt tokens at block 0. The two arms instantiate either one shared classifier or a classifier per domain, holding the nominal backbone, prompt design, stream, optimisation settings, and seed schedule fixed. The shared and per-domain arms are separately trained banks. This intervention changes a training design; it need not preserve predictive quality. It is also distinct from changing a component pairing at a single frozen checkpoint.

The recorded runs use 20 epochs per domain on Office-Home and 5 on DomainNet, with learning rate 0.01. The training routine uses AdamW, zero weight decay, and no learning-rate schedule; the runner's batch-size setting is 64. Only the active domain prompt and the corresponding trainable classifier parameters are updated, with the backbone frozen. The shared classifier continues across domains; domain-specific classifiers are independently instantiated. The recorded selector fits five K-means centroids per domain and uses the layer-12 prompt-free query with the maximum cosine similarity to a domain's keys. The same nominal optimisation recipe is used within each paired classifier-scope contrast.

Retention and update budget in the shared arm. The shared classifier is updated over the whole stream with no rehearsal buffer, while each per-domain head is fit once and never overwritten, so the contrast could in principle measure retention rather than readout scope. Table 14 checks a simple recency pattern: for each domain it reports how well that domain's own branch classifies that domain's test images, in both arms. On Office-Home the per-domain advantage falls with position in the stream, apart from a rise at Clipart, from $+6.31$ pp on the first domain to -0.59 pp on the last, with a correlation of -0.82 between the gap and stream position that is negative in all three seeds. That is the signature a retention account predicts, and we therefore do not read the Office-Home contrast as isolating readout scope. DomainNet has no consistent monotone recency pattern: the correlation is -0.20 and changes sign across seeds, and the largest gap belongs to quickdraw at position three rather than to an early domain. This does not exclude retention or interference effects. The pooled DomainNet contrast is a difference between trained banks under identical assignments.

The arms share the per-domain training loop and epoch count, but each per-domain head is fitted only on its own domain whereas the shared head is updated throughout the stream. These different optimisation histories can affect convergence and retention, with no guaranteed direction for the correction-value contrast. This comparison therefore identifies a difference between trained banks under identical assignments, not an isolated effect of classifier scope.

Table 14: **Does the shared classifier retain earlier domains?** Own-domain accuracy in percent: branch d evaluated on the test images whose true domain is d, averaged over three seeds. Gap is per-domain minus shared. A retention account predicts the gap to shrink with stream position. The correlation between gap and position is -0.82 on Office-Home (per seed -0.81, -0.72, -0.92) and -0.20 on DomainNet (per seed -0.16, $+0.14$, -0.39).

Data	Position	Domain	Shared	Per-domain	Gap
OH	0	Art	77.84	84.16	+6.31
OH	1	Clipart	69.17	78.24	+9.07
OH	2	Product	91.26	95.05	+3.79
OH	3	Real World	90.98	90.39	−0.59
DN	0	clipart	55.40	67.52	+12.12
DN	1	infograph	18.41	15.50	−2.92
DN	2	painting	51.44	65.52	+14.08
DN	3	quickdraw	4.28	41.37	+37.09
DN	4	real	73.33	81.46	+8.14
DN	5	sketch	52.51	40.97	−11.54

The substrate's recorded environment resolves `timm` (Wightman, 2019) 0.9.8's pretrained `vit_base_patch16_224` to `augreg2_in21k_ft_in1k`. This differs from the separate AugReg weights used in S-iPrompts. The constructor name alone is therefore insufficient to establish checkpoint equivalence. Our cross-system comparisons do not treat the substrate as an architecture-matched control for S-iPrompts, or as an additional published method.

For each paired seed, we apply the shared arm's recorded routing assignments to both trained banks, and separately repeat with the per-domain arm's assignments. Within a common-reference comparison, each sample has exactly the same selected domain index and the same mismatch membership in both banks. The two references reuse the same trained pair. They remove assignment differences; they do not match training outcomes or identify a unique microscopic mechanism.

Table 15: **Classifier scope by domain.** Domain-level classifier-scope contrasts. The first three numerical columns use the common shared-arm router; entries are per-domain-bank minus shared-bank changes in repair contribution, harm contribution, and net correction value. The final column reports net change under the common per-domain-arm router. Domain rows use their own N_d; pooled rows use full test N. Values are mean and sample SD in pp.

		Common shared-arm assignments			Other common reference
Data	Domain	Δ repair	Δ harm	Δ net	Δ net
OH	Art	4.80 ± 0.50	1.45 ± 0.80	3.35 ± 1.04	3.88 ± 0.99
	Clipart	2.65 ± 0.28	0.56 ± 0.81	2.09 ± 0.75	2.09 ± 0.47
	Product	3.46 ± 0.46	0.15 ± 0.13	3.32 ± 0.33	2.95 ± 0.66
	Real World	2.93 ± 0.50	1.22 ± 0.67	1.71 ± 1.03	1.34 ± 1.16
	Pooled	3.30 ± 0.17	0.77 ± 0.30	2.53 ± 0.38	2.41 ± 0.35
DN	clipart	6.24 ± 2.26	0.41 ± 0.55	5.83 ± 2.69	5.95 ± 2.60
	infograph	0.64 ± 0.99	2.31 ± 1.84	-1.67 ± 2.82	-1.67 ± 2.69
	painting	6.09 ± 0.67	0.71 ± 0.34	5.39 ± 1.01	5.33 ± 1.09
	quickdraw	0.71 ± 0.10	-0.17 ± 0.06	0.88 ± 0.13	0.88 ± 0.13
	real	4.69 ± 0.99	0.23 ± 0.17	4.47 ± 1.11	4.51 ± 1.10
	sketch	2.72 ± 2.60	1.81 ± 3.99	0.91 ± 6.58	0.99 ± 6.63
	Pooled	3.24 ± 0.24	0.56 ± 0.40	2.68 ± 0.21	2.71 ± 0.23

The full-population correction-value contrast is positive in every paired seed for both common references on both benchmarks. Under the shared-arm reference, repairs and harms both increase in every pooled seed, with the repair increase larger. Table 15 shows why this pooled pattern should not be asserted for every domain: infograph has a negative mean net contrast, sketch varies con-

siderably, and quickdraw's harm contribution decreases. The accompanying CSV files preserve every seed/domain integer count and both reference routers. Weighted domain counts reconstruct the pooled contrasts exactly.

E.1 PER-DOMAIN ACCOUNTING ON THE PRIMARY BANKS

Table 16 repeats the four-cell accounting of Table 2 for each DomainNet domain of the S-iPrompts banks, with the native router and each domain's own test images as the denominator; quickdraw and real hold 59% of the test set, so a pooled effect could in principle be carried by those two domains. It is not. The prompt-only action is negative on four of the six domains, positive only on painting ($+0.26$ pp) and within 0.01 pp of zero on quickdraw, whose router is wrong on fewer than 1% of its images; the head-only action exceeds the prompt-only action on every domain; and the domain-mean values, -0.201 ± 0.044, $+3.808 \pm 0.035$, and $+4.476 \pm 0.064$ pp, give the head-only action 85% of the joint gain against 89% pooled, with the prompt-only action negative in every seed under either average.

Table 16: **Four-cell actions per DomainNet domain, S-iPrompts.** Native router; pp of each domain's test images; mean $\pm$ sample SD over three seeds. The pooled row is the DomainNet column of Table 2.

Domain	Test images	Prompt only	Head only	Joint	I_{PH}
clipart	14,604	-0.160 ± 0.156	$+4.862 \pm 0.428$	$+6.126 \pm 0.473$	$+1.424 \pm 0.141$
infograph	15,582	-0.075 ± 0.052	$+1.493 \pm 0.412$	$+1.848 \pm 0.079$	$+0.430 \pm 0.326$
painting	21,850	$+0.261 \pm 0.167$	$+5.494 \pm 0.233$	$+6.673 \pm 0.089$	$+0.918 \pm 0.223$
quickdraw	51,750	$+0.005 \pm 0.006$	$+0.175 \pm 0.061$	$+0.263 \pm 0.069$	$+0.084 \pm 0.065$
real	52,041	-0.930 ± 0.231	$+6.385 \pm 0.110$	$+6.563 \pm 0.085$	$+1.108 \pm 0.271$
sketch	20,916	-0.309 ± 0.168	$+4.438 \pm 0.040$	$+5.385 \pm 0.189$	$+1.256 \pm 0.109$
Domain mean	–	-0.201 ± 0.044	$+3.808 \pm 0.035$	$+4.476 \pm 0.064$	$+0.870 \pm 0.103$
Pooled	176,743	-0.297 ± 0.064	$+3.669 \pm 0.039$	$+4.141 \pm 0.038$	$+0.769 \pm 0.117$

F MATCHED CLASSIFIER-EXPOSURE INTERVENTION

We reuse the twelve completed per-domain-head banks: S-iPrompts and CP-Prompt, each on Office-Home and DomainNet with training seeds 87, 1013, and 1993. The visual backbone, every prompt, the original routing assignments, and the test images remain fixed. For each domain head, both refitting arms start from the same native head and use the same domain's fit images, labels, image order, and number of updates. Own always encodes a fit image with that head's paired domain prompt. Mixed exposure encodes half of the images in each epoch with other prompts: $\lfloor N_{\mathrm{fit},j}/2 \rfloor$ assignments are distributed over the $T-1$ foreign prompts with counts differing by at most one; the other assignments use the paired prompt. This controls image exposure and update count without multiplying the training budget by the number of prompts. Fit, validation, and test features are cached once using deterministic native evaluation preprocessing; no new stochastic image augmentation is sampled during refitting.

What is refitted in each family. In S-iPrompts only the native linear heads are refitted, using cross-entropy, SGD with momentum 0.9, batch size 128, and cosine learning-rate decay; weight decay is 2×10^{-4}, except for the first DomainNet head, which retains 5×10^{-4}. In CP-Prompt the classifier is not a linear head but a set of class prototypes generated by a frozen text encoder from learnable text-context parameters (`ctx`); the refit updates only those native text-context parameters, keeping the image-side common prompt and personalised prefixes, the text encoder, and the native feature normalisation fixed. No surrogate linear head is substituted, so the refitted arms remain the native predictor with a different classifier exposure. Both families use the same predefined grid of learning rates $\{0.001, 0.01\}$ and durations $\{5, 10\}$ epochs. Within a trained bank, one configuration is chosen for both arms by the arithmetic mean of their validation accuracies under the original routing assignments, locked before any test evaluation; ties prefer fewer epochs and then the smaller learning rate. The selected learning rate is 0.001 for all twelve banks. Selected durations for seeds

$87/1013/1993$ are $5/5/10$ epochs for S-iPrompts on Office-Home and $10/10/10$ on DomainNet; for CP-Prompt they are $5/5/5$ on Office-Home and $10/5/5$ on DomainNet.

Pre-specification. The study plan fixed, before any arm was evaluated on test data (Cawley & Talbot, 2010): the primary contrasts, $I_{PH,\text{Mixed}} - I_{PH,\text{Own}}$ and $H_{\text{Mixed}} - H_{\text{Own}}$ on the three S-iPrompts DomainNet banks; the performance-preservation contrasts $A_{ss,\text{Mixed}} - A_{ss,\text{Own}}$ and $A_{dd,\text{Mixed}} - A_{dd,\text{Own}}$ with a 0.5 full-test pp non-inferiority reference; the 50% uniform foreign-prompt rule; the learning-rate/epoch grid; the selection rule (mean of the two arms' validation current accuracy, ties to fewer updates then the smaller learning rate); and that Own is reported against the native checkpoint separately. Against that reference, ΔA_{ss} and ΔA_{dd} are -0.108 and -0.108 (S-iPrompts DN), -0.114 and -0.041 (S-iPrompts OH), and -0.114 and -0.062 pp (CP-Prompt OH), all inside the margin; only CP-Prompt DomainNet, at -1.023 and -0.877 pp, exceeds it. CP-Prompt was designated the secondary family, whose near-zero native interaction was noted as a boundary rather than an automatic replication. Test evaluation followed the locked selection.

Validation partitions. Office-Home validation consists of a deterministic, domain/class-stratified approximately 10% partition of the original training images, held out from refitting; these images were available when the frozen native bank was originally trained. DomainNet uses the existing external validation split, excluded from the original bank fitting manifests. Both arms use the same partition. The experiment revisits all domains' fit data after stream training and is a post-training diagnostic, not a rehearsal-free continual-learning method.

Contrasts and uncertainty. Within each fitted arm we evaluate the same four assignments $(s, s), (d, s), (s, d), (d, d)$ without rerouting. The reported arm contrast is Mixed minus Own. Interaction is computed from the four correctness indicators for each image before aggregation. Prompt-switch harm in each arm is the fraction of all test images whose prediction under that arm's (s, s) changes from correct to incorrect under (d, s). Thus the assignments and scored population are common across arms, while the two arms need not have identical baseline-correct sets. Own versus the original native checkpoint is reported here rather than inside the exposure contrast, so that continued classifier fitting is not attributed to mixing. Own $-$ native is $+0.617 \pm 0.202$ pp in A_{ss} and $+0.630 \pm 0.199$ pp in A_{dd} on the S-iPrompts DomainNet banks, $+0.135 \pm 0.025$ and $+0.133 \pm 0.013$ pp in CP-Prompt DomainNet, $+0.114 \pm 0.532$ and $+0.187 \pm 0.431$ pp in S-iPrompts Office-Home, and $+0.114 \pm 0.160$ and $+0.010 \pm 0.048$ pp in CP-Prompt Office-Home, so neither arm is measured against a degraded baseline. Refitting alone changes I_{PH} by -0.049 ± 0.022 pp on the primary DomainNet contrast, against -0.398 pp for mixing. We report the mean and sample SD of the three within-bank arm contrasts. Separate 95% paired bootstrap intervals use $5{,}000$ resamples of images within each domain, applying each sampled image index to both arms and all three fixed banks. These intervals quantify test-image uncertainty conditional on the fitted banks and selected configurations; they do not include training-seed or configuration-selection uncertainty. Training seeds and images are not pooled as independent replicates.

Control arms. Two additional refits reuse the cached fit features, the locked configuration, the native head initialisation, and the seed of the completed run; nothing else changes. *Own-half* keeps, in every epoch, exactly the slots that Mixed assigns to the head's own prompt and drops the foreign slots, either as they are (half of Mixed's updates) or processed twice (the same update count); it therefore matches Mixed's own-prompt data with no foreign features. *Prompt-free* refits each head on the prompt-free features of its own fit images with the Own budget; its test accuracy under prompt-free features, against the Own bank's A_{dd} and A_{ss}, is the removal cost that survives refitting the head without prompts. Table 17 reports both Own-half budgets against Own, next to Mixed; Table 18 reports the Prompt-free refit.

Interpretation of the completed intervention. In S-iPrompts on DomainNet, Mixed exposure reduces interaction and prompt-switch harm in every seed, but also reduces repair and costs approximately 0.108 pp in both original-routing and domain-diagonal accuracy relative to Own. Prompt replacement remains harmful under the selected head and helpful under the domain head; in CP-Prompt it is helpful under both heads in both arms. The intervention therefore partially changes the switch effects rather than eliminating head dependence or establishing a performance-improving policy. S-iPrompts' Office-Home interaction and harm contrasts are smaller and vary in sign across

Table 17: **Exposure versus dilution.** Change relative to the Own arm, full-test pp, paired three-seed mean ± sample SD. Own-half sees Mixed's own-prompt slots and nothing else.

Family	Data	Arm	ΔI_{PH}	ΔA_{ss}
S-iPrompts	DN	Mixed (foreign features, all slots)	-0.398 ± 0.053	-0.108 ± 0.046
S-iPrompts	DN	Own-half, dropped slots (half the updates)	$+0.012 \pm 0.014$	-0.083 ± 0.036
S-iPrompts	DN	Own-half, slots twice (same updates)	$+0.002 \pm 0.011$	-0.032 ± 0.018
CP-Prompt	DN	Mixed (foreign features, all slots)	-0.351 ± 0.106	-1.023 ± 0.073
CP-Prompt	DN	Own-half, dropped slots (half the updates)	-0.005 ± 0.016	-0.027 ± 0.030
CP-Prompt	DN	Own-half, slots twice (same updates)	$+0.002 \pm 0.024$	-0.017 ± 0.012
S-iPrompts	OH	Mixed (foreign features, all slots)	-0.207 ± 0.406	-0.114 ± 0.177
S-iPrompts	OH	Own-half, dropped slots (half the updates)	$+0.145 \pm 0.212$	-0.000 ± 0.328
S-iPrompts	OH	Own-half, slots twice (same updates)	-0.041 ± 0.065	$+0.124 \pm 0.224$
CP-Prompt	OH	Mixed (foreign features, all slots)	-0.031 ± 0.194	-0.114 ± 0.252
CP-Prompt	OH	Own-half, dropped slots (half the updates)	-0.073 ± 0.048	-0.052 ± 0.048
CP-Prompt	OH	Own-half, slots twice (same updates)	-0.021 ± 0.126	-0.114 ± 0.095

Table 18: **Removal after retraining the head.** Full-test pp, three-seed mean ± sample SD. Column 1: prompt-free features read by the Own-refitted domain heads without further refitting, against the Own bank's A_{dd}. Columns 2–3: prompt-free features read by heads refitted on prompt-free features, the domain head against the Own bank's A_{dd} and the selected head against its A_{ss}. Paired within bank; SD over three seeds.

Family	Data	No refit	Refit vs A_{dd}	Refit vs A_{ss}
S-iPrompts	DN	-25.327 ± 1.244	-10.451 ± 1.130	-10.104 ± 1.132
CP-Prompt	DN	-16.552 ± 0.106	-11.441 ± 0.656	-10.683 ± 0.547
S-iPrompts	OH	-4.710 ± 1.303	-2.863 ± 0.925	-2.905 ± 0.845
CP-Prompt	OH	-4.928 ± 0.144	-4.036 ± 0.144	-3.351 ± 0.252

seeds. CP-Prompt reproduces the signed direction of the interaction change in every DomainNet seed at a larger accuracy cost, while its interaction crosses zero rather than shrinking (Table 19); the two families therefore share the direction, not the magnitude or the mechanism. Neither family identifies a unique cause of the original training dynamics. DualPrompt+PDH follows the S-iPrompts pattern within the accuracy margin (Appendix H).

Table 19: **Signed change and magnitude change are different readings.** Arm-level prompt–head interaction in full-test pp, averaged over three seeds. The absolute columns are the magnitude of the three-seed mean, not the mean of the per-seed magnitudes. A signed decrease can enlarge the magnitude when the value crosses zero, so the signed and absolute columns support different words.

Family	Data	I_{PH} Own	I_{PH} Mixed	$\lvert I_{PH}\rvert$ Own	$\lvert I_{PH}\rvert$ Mixed	Magnitude
S-iPrompts	OH	$+0.311$	$+0.104$	0.311	0.104	decreases
S-iPrompts	DN	$+0.719$	$+0.321$	0.719	0.321	decreases
CP-Prompt	OH	-0.052	-0.083	0.052	0.083	increases
CP-Prompt	DN	$+0.145$	-0.206	0.145	0.206	increases
DualPrompt+PDH	DN	$+0.308$	$+0.156$	0.308	0.156	decreases

On DomainNet, the interval for the average interaction contrast is $[-0.438, -0.358]$ pp and for harm is $[-0.209, -0.165]$ pp. Current accuracy under the selected head, ΔA_{ss}, also falls, with interval $[-0.153, -0.063]$ pp. The Office-Home intervals include zero; neither this uncertainty nor the sign variation supports a consistent replication of the DomainNet effect. These intervals are pointwise and are not corrected for simultaneous inference. The per-image contrast is first averaged over the three fixed banks, then resampled within domains with domain sample sizes preserved; seed–image rows are never treated as separate images.

Common-baseline sensitivity check. For each seed, partition test images by their correctness under the Own and Mixed original-assignment predictors: correct under both, lost correctness, gained correctness, or wrong under both. In each arm define the harm indicator $H_i = \mathbf{1}[c_{ss,i} =$

Table 20: **Seed SDs and image intervals.** Exposure contrasts with two distinct uncertainty summaries. Mixed minus Own, full-test pp. SD describes three training seeds. Pointwise 95% paired image-bootstrap intervals condition on those banks and the validation-selected head fits; they do not include training or model-selection uncertainty.

Family	Data	Contrast	Mean ± SD	Image 95% CI
S-iPrompts	OH	ΔI_{PH}	-0.207 ± 0.406	$[-0.425, +0.010]$
S-iPrompts	OH	Δ harm	-0.104 ± 0.140	$[-0.228, +0.021]$
S-iPrompts	OH	ΔA_{ss}	-0.114 ± 0.177	$[-0.342, +0.104]$
S-iPrompts	OH	ΔA_{dd}	-0.041 ± 0.234	$[-0.259, +0.176]$
S-iPrompts	DN	ΔI_{PH}	-0.398 ± 0.053	$[-0.438, -0.358]$
S-iPrompts	DN	Δ harm	-0.187 ± 0.015	$[-0.209, -0.165]$
S-iPrompts	DN	ΔA_{ss}	-0.108 ± 0.046	$[-0.153, -0.063]$
S-iPrompts	DN	ΔA_{dd}	-0.108 ± 0.023	$[-0.154, -0.064]$
CP-Prompt	OH	ΔI_{PH}	-0.031 ± 0.194	$[-0.218, +0.145]$
CP-Prompt	OH	Δ harm	-0.010 ± 0.065	$[-0.124, +0.093]$
CP-Prompt	OH	ΔA_{ss}	-0.114 ± 0.252	$[-0.342, +0.094]$
CP-Prompt	OH	ΔA_{dd}	-0.062 ± 0.300	$[-0.259, +0.135]$
CP-Prompt	DN	ΔI_{PH}	-0.351 ± 0.106	$[-0.391, -0.312]$
CP-Prompt	DN	Δ harm	-0.089 ± 0.036	$[-0.104, -0.073]$
CP-Prompt	DN	ΔA_{ss}	-1.023 ± 0.073	$[-1.080, -0.963]$
CP-Prompt	DN	ΔA_{dd}	-0.877 ± 0.074	$[-0.935, -0.819]$
DualPrompt+PDH	DN	ΔI_{PH}	-0.152 ± 0.029	$[-0.182, -0.123]$
DualPrompt+PDH	DN	Δ harm	-0.068 ± 0.012	$[-0.083, -0.053]$
DualPrompt+PDH	DN	ΔA_{ss}	-0.319 ± 0.006	$[-0.357, -0.280]$
DualPrompt+PDH	DN	ΔA_{dd}	-0.290 ± 0.016	$[-0.328, -0.251]$

$1, c_{ds,i} = 0]$ and sum $H_i^{\text{Mixed}} - H_i^{\text{Own}}$ over each partition, dividing every contribution by the same full N. The partition's contributions sum exactly to the reported harm contrast. The

Table 21: **Harm partitioned by baseline correctness.** Harm-contrast sensitivity to baseline correctness. Mixed minus Own contributions in full-test pp, paired three-seed mean ± SD. The subsets are defined jointly from both fitted arms and vary across seeds. Both-wrong inputs contribute zero harm. Contributions sum within each seed; SDs do not add.

Family	Data	Correct in both	Lost correctness	Gained correctness	Total
S-iPrompts	OH	-0.156 ± 0.062	-0.052 ± 0.036	$+0.104 \pm 0.048$	-0.104 ± 0.140
S-iPrompts	DN	-0.236 ± 0.020	-0.068 ± 0.002	$+0.117 \pm 0.016$	-0.187 ± 0.015
CP-Prompt	OH	-0.062 ± 0.031	$+0.000 \pm 0.000$	$+0.052 \pm 0.036$	-0.010 ± 0.065
CP-Prompt	DN	-0.085 ± 0.034	-0.057 ± 0.007	$+0.053 \pm 0.001$	-0.089 ± 0.036

common-correct contribution is negative in all three seeds on both benchmarks and in both families. This does not turn the inconsistent pooled OH contrast into a replication: the changed baseline-correct sets contribute in the opposite direction. Nor is it an estimate for a population defined before fitting. For images wrong under both original-assignment predictors, the repair contrast contributes -0.082 ± 0.030 full-test pp on DN and 0.000 ± 0.082 pp on OH in S-iPrompts, and -0.160 ± 0.025 pp on DN and -0.041 ± 0.048 pp on OH in CP-Prompt. Complete per-seed integer partitions and the recovery of total repair, harm, and baseline-accuracy differences are supplied in `data/exposure/common_correct_audit.json` (S-iPrompts) and `data/exposure/common_correct_two_families.json` (both families, recomputed from the stored per-image arm predictions; the S-iPrompts rows reproduce the audit file).

F.1 TIED-PROMPT BANKS

Three tied-prompt DomainNet banks, referred to in Section 5, retrain the native S-iPrompts recipe with one change: the six prompt entries are tied to a single parameter that is unfrozen at every stage and so trained through the whole stream, while each per-domain head is trained at its own stage and frozen afterwards, as in the native banks. Initialisation, KMeans keys, optimiser, schedule, seeds $(87/1013/1993)$, data order, and epochs are native, and a per-stage guard checks that the trainable

set is exactly the current head and the tied prompt and that every other head is bitwise unchanged. The router reads prompt-free features, so routing accuracy and $\mathcal{M}$ coincide with the native banks image by image, and evaluation and the Own refit use the native cache and the locked grid above without modification. Every head except the last was fitted to features of a prompt that continued to change. A classifier-only refit tests whether the loss persists after adapting the head to the final tied prompt; it does not isolate all possible causes of that loss.

Completed results and their reference. Before refitting, the tied-prompt banks lose 4.67 ± 2.00 pp of routed accuracy relative to the native untied banks; 89% of the pooled loss comes from quickdraw. The joint domain-correction gain changes by -0.251 ± 0.040 pp. This contrast measures a correction, not the overall quality of the retrained bank. After Own refitting, tied and untied routed accuracies are $62.25 \pm 0.09\%$ and $62.44 \pm 0.92\%$, respectively: the paired difference is -0.19 ± 1.01 pp and changes sign across seeds. Thus refitting reduces the mean deficit to 0.19 pp. Removing the tied prompt still costs 19.20 ± 0.79 pp. This follow-up is post-hoc. With the same prompt in every slot, prompt-only domain correction and I_{PH} are zero by construction, regardless of these accuracy differences.

The domain pattern is consistent with a compatibility account: the loss concentrates on quickdraw, whose native removal cost is the largest of the six (50 pp), whereas sketch, the last domain and hence the head fitted to the final prompt, improves in every seed. This pattern and the refit response support examining compatibility between the final prompt and the heads trained earlier in the stream.

G Further seeds and a reversed domain order

The S-iPrompts DomainNet recipe of Appendix A was run again for two further training seeds, 7 and 42, and three more times with the domain order reversed, sketch, real, quickdraw, painting, infograph, clipart, for seeds 87, 1013 and 1993. Nothing else changed: the same images and splits, the same optimiser, schedule and epochs, the same nearest-centroid router, and the same per-stage audits. A bank trained in the reversed order is its own bank, since its stage index names a different domain, so the two orders are reported separately and are never pooled.

Table 22 reads the four cells of Table 2 on all eight banks. The prompt-only action is negative in every one of the five canonical-order banks and in every reversed one; the head-only action exceeds it everywhere; and the head-only / joint contrast ratio is 88.7% over five seeds against 88.6% on the three primary banks, the 89% of Section 4.1, and 87.4% when the order is reversed. Domain-routing accuracy is $80.46 \pm 0.09\%$ over the five canonical banks and $80.38 \pm 0.11\%$ over the reversed ones.

Table 22: **The decomposition on five canonical-order banks and three reversed-order banks.** Full-test pp against the native selected-prompt/selected-head prediction A_{ss}; † marks a bank that does not enter the main tables. The two orders are summarised separately.

Order	Seed	Prompt only	Head only	Joint	I_{PH}
Canonical order	87	-0.238	$+3.704$	$+4.100$	$+0.634$
	1013	-0.287	$+3.628$	$+4.176$	$+0.835$
	1993	-0.366	$+3.675$	$+4.146$	$+0.837$
	7†	-0.234	$+3.722$	$+4.203$	$+0.715$
	42†	-0.404	$+3.588$	$+4.021$	$+0.836$
Mean ($n = 5$)		-0.306 ± 0.076	$+3.664 \pm 0.055$	$+4.129 \pm 0.072$	$+0.771 \pm 0.093$
head-only / joint		88.7%			
Reversed order	87†	-0.270	$+3.547$	$+4.165$	$+0.888$
	1013†	-0.284	$+3.790$	$+4.199$	$+0.693$
	1993†	-0.278	$+3.590$	$+4.137$	$+0.826$
Mean ($n = 3$)		-0.277 ± 0.007	$+3.642 \pm 0.130$	$+4.167 \pm 0.031$	$+0.802 \pm 0.100$
head-only / joint		87.4%			

The matched classifier-exposure intervention of Appendix F was repeated on the two further canonical banks and on the three reversed banks with the locked grid and selection rule (Tables 23

and 24). The signed interaction decreases in all five canonical banks, -0.395 ± 0.038 pp against the -0.398 ± 0.053 pp of the three primary ones, and their Own-half and prompt-free controls hold as well: refitting on the same own-prompt slots without the foreign features moves the interaction by $+0.007 \pm 0.016$ pp at half the updates and by -0.003 ± 0.016 pp at the same update count, and a prompt-free refit still leaves 10.25 ± 0.83 pp of removal cost against the Own arm's A_{ss} under the selected head. It decreases in all three reversed banks too, by 0.397 ± 0.107 pp, while their paired joint-gain change is -0.008 ± 0.010 pp. The agreement between the two means is closer than the seeds warrant: the per-seed reversed values span -0.480 to -0.276 against -0.434 to -0.338, so what replicates is the direction in every bank and the overlap of the two ranges, not the coincidence of the two averages.

Table 23: **Classifier exposure on the five canonical-order banks.** Mixed minus Own, full-test pp; † marks a bank that does not enter the main tables, which report the three primary banks (Table 3). The two diagonal cells move together, which is why ΔJ stays near zero while ΔI_{PH} does not.

Seed	ΔI_{PH}	Δ harm	ΔA_{ss}	ΔA_{dd}	ΔJ
87	-0.338	-0.187	-0.147	-0.101	$+0.046$
1013	-0.423	-0.172	-0.057	-0.090	-0.033
1993	-0.434	-0.201	-0.119	-0.135	-0.015
7†	-0.382	-0.158	-0.207	-0.214	-0.007
42†	-0.397	-0.166	-0.143	-0.166	-0.023
Mean	-0.395 ± 0.038	-0.177 ± 0.017	-0.135 ± 0.054	-0.141 ± 0.051	-0.007 ± 0.031

Table 24: **Classifier exposure on the reversed-order banks.** Mixed minus Own, full-test pp, three seeds. The canonical-order values are in Table 23.

Seed	ΔI_{PH}	Δ harm	ΔA_{ss}	ΔA_{dd}	ΔJ
87	-0.433	-0.155	-0.160	-0.165	-0.005
1013	-0.276	-0.105	-0.191	-0.210	-0.019
1993	-0.480	-0.162	-0.105	-0.105	-0.001
Mean	-0.397 ± 0.107	-0.141 ± 0.031	-0.152 ± 0.044	-0.160 ± 0.053	-0.008 ± 0.010

Two boundaries hold this section to what it tested. A single reversal is one more order, not a sample of the 720 possible ones: the pattern persists under this reversed curriculum, without establishing order invariance. And the five new banks carry both the decomposition and the exposure contrast; the alternate-router contrasts of Appendix D use the three primary banks, with only the protocol-locked Gaussian router also reported on the two added canonical-order banks, and the Office-Home settings were not re-run, so the main tables keep the three primary banks throughout.

For each reversed bank the four cells were computed twice, once from the cached features used here and once from fresh native predictions extracted before classifier refitting. These paths share model and dataset infrastructure; they are separate extractions, not independent implementations. The reconstructed four accuracies agree within 0.001 pp, below the gate's 0.01 pp tolerance. Recorded order and domain-count checks support the stage-to-domain mapping. The agreement verifies aggregate numerical consistency, not image identity on its own. The source package includes the gate records and integer arm counts.

H DUALPROMPT+PDH: AN EVALUATION ADAPTATION WITH A PER-DOMAIN HEAD BANK

System and modifications. DualPrompt comes from the mammoth implementation lineage; its import path contains no S-Prompts file (Wang et al., 2022b). Its general prompt (blocks 0–1) is shared across domains. Its expert prompt, a prefix at blocks 2–4, is specific to each domain and is selected by key–query matching. The released model has one shared 345-way classifier. We made five changes to the model, each recorded in every saved artifact:

1. a prefix-reshape fix, so that key and value prefixes are no longer mixed across images in a batch;
2. a bank of six identically initialised linear heads, where stage t trains head t together with expert prompt t and the shared general prompt, and all earlier heads, expert prompts, and keys stay bitwise frozen;
3. a pool of six expert prompts with top-1 selection, so that the prompt index equals the domain index;
4. per-image rather than batch-wise prompt selection at test time;
5. no validation-based checkpoint selection, so each stage keeps its last epoch.

Everything else follows our existing DomainNet recipe for this code base: Adam with learning rate 10^{-4}, batch size 64, and 20 epochs per stage on a ViT-B/16 pretrained on ImageNet-21k. The pretrained weights and preprocessing differ from those of S-iPrompts, so we do not compare accuracy levels across families. Training used TF32 arithmetic; feature extraction, exposure refits, and every reported cell use full float32 precision.

Pre-specification and gates. The criteria were written and hashed before the first training step: the head-only action exceeds the prompt-only action, the Own-arm interaction is positive, and the Mixed-minus-Own interaction change is negative, each in all three seeds, with a mean change of at least 0.10 pp in magnitude. Three gates had to pass first: code tests with a smoke run, a survival check after stage 3, and a completed-bank check (routed accuracy at least 40%, domain identification at least one third, no prompt serving more than 60% of images, and a head-only action of at least 0.5 pp). All gates passed, and each criterion then held in every seed. No seed was re-run. Table 25 gives the per-seed cells.

Table 25: **DualPrompt+PDH DomainNet cells per seed.** Fixed adapted banks, full-test pp except A_{ss} (%). Removal drops the prompts under the selected head. The last two columns give exact McNemar p-values for prompt correction under the selected and the domain head.

Seed	A_{ss}	$A_{ds} - A_{ss}$	$A_{dd} - A_{sd}$	Head only	I_{PH}	Removal	p (sel.)	p (dom.)
87	53.24	−0.055	+0.292	+3.794	+0.347	−19.20	4.7e-02	4.3e-22
1013	53.11	−0.142	+0.155	+3.907	+0.297	−19.42	5.9e-07	3.1e-07
1993	52.76	−0.144	+0.200	+3.833	+0.345	−18.95	5.2e-07	5.0e-11

Neighbouring expert prompts are similar. Each stage initialises its expert prompt from the previous one, so neighbouring prompts stay close. Across seeds, the mean off-diagonal cosine similarity between the six trained expert prompts is 0.91, and 4 of the five neighbouring pairs exceed 0.95 (maximum 0.987; Table 26). This high similarity is consistent with DualPrompt+PDH's small prompt-only magnitudes. For this reason the criteria were written on the head action and the interaction change rather than on the size of the prompt-only action. The sign reversal itself holds in every seed (Table 25).

Table 26: **Cosine similarity of neighbouring DualPrompt expert prompts.** Pair $(t, t{+}1)$ compares the prompts of consecutive domains in training order; the last column averages all 30 off-diagonal pairs.

Seed	$(1,2)$	$(2,3)$	$(3,4)$	$(4,5)$	$(5,6)$	Mean
87	0.96	0.97	0.93	0.98	0.99	0.909
1013	0.96	0.97	0.93	0.98	0.99	0.910
1993	0.96	0.97	0.93	0.98	0.99	0.910

Exposure. The Own and Mixed refits of the six linear heads use the same fork of the exposure code and the same validation-based configuration rule as the other configurations (Appendix F). Own alone raises routed accuracy by $+1.25 \pm 0.11$ pp over the fixed DualPrompt+PDH bank. Mixed then reduces the interaction in every seed, from 0.308 to 0.156 pp on average. Under Mixed,

prompt correction changes accuracy by -0.037 pp under the selected head and $+0.119$ pp under the domain head, and the sign reversal remains in 2 of three seeds. Tables 19 and 20 include the DualPrompt+PDH rows. We did not run own-half controls for this configuration.

I BOUNDARIES OF THE EVIDENCE

Interpretive limits and reusable evidence. The experiments distinguish quantities that can otherwise be conflated, but do not establish a new expressive-capacity bound, a universal requirement for per-domain classifiers, or where knowledge is stored during training. The head-conditioned reversal and the consistent exposure response occur on DomainNet, in S-iPrompts and DualPrompt+PDH; Office-Home provides boundaries rather than uniform confirmations; CPrompt shows the same large removal cost and sub-pp correction effect on DomainNet under a combined readout, but has no head factor and no exposure arm. The primary design uses three seeds and one order per setting. The additional seeds and reversed-order S-iPrompts DomainNet banks (Appendix G) extend that recipe only; broad robustness to pretraining, architecture, stream order, or future tasks is untested. The exhaustive oracle is descriptive and uses labels; obtaining its opportunity with a practical selector remains a separate empirical question.

The source bundle contains the completed component-interchange and policy-comparison summaries, paired component counts, common-router counts and contrasts, CPrompt removal/correction counts, and the displayed-evidence export used to generate the tables. Exact expectation rows are marked separately from integer-count interventions. This makes this paper's numerical claims traceable without incorporating unfinished experiments or treating an audit of internal consistency as a substitute for independent model replication.

J ADDITIONAL CONTEXT AND RELATION TO PRIOR WORK

Prompt selection in continual learning. L2P, native DualPrompt, and CODA-Prompt select or compose prompts under a shared readout, so changing that selection or composition retains the classifier by construction (Wang et al., 2022c;b; Smith et al., 2023). S-Prompts retains independent domain prompts and classifiers, while PINA and SOYO study parameter selection in DIL (Wang et al., 2022a; 2024; 2025b). Improving selection is not every method's primary contribution; selection is nevertheless part of their inference rule. In a banked system, a prompt-only deployment also arises when the selector or prompt bank changes while classifier assignments are retained. Our protocol does not require a four-cell prompt–head crossing for shared-readout methods. Joint metrics likewise remain appropriate when deployment switches the full parameter branch; they simply do not identify the effect of changing the prompt alone. Our contrast is between the router's own assignment and the true-domain assignment under a retained classifier. It is not a comparison against random or fixed prompt selection, which changes a far larger share of the assignments and therefore answers a different question.

Task inference and component-level interventions. HiDe-Prompt and HiDe-PET distinguish task inference from within-task prediction (Wang et al., 2023; 2025a). Kim et al. factor class probabilities into within-task and task-identity terms and analyse corresponding cross-entropy bounds (Kim et al., 2022). These perspectives clarify why task-identification accuracy alone need not describe final classification. Our four-cell evaluation instead holds a trained DIL bank fixed and separately changes its prompt and classifier assignments. It identifies the effect of each specified action on the same images; the matched exposure study then tests how classifier training changes those effects.

Consistency and alternatives to prompt selection. CPrompt's consistency objective and CPS-Prompt's separate prompt/head fitting make classifier training a design variable (Gao et al., 2024; Lim et al., 2026). SinglePrompt questions prompt selection in task-free online continual learning (Park et al., 2026); other work examines query-based selection or strong prototype and random-projection readouts on frozen features (Thede et al., 2024; Zhou et al., 2025; McDonnell et al., 2023). These methods operate under different settings and training interventions. Our paired substitutions distinguish reliance on a trained prompt from the value of correcting its domain identity, while the

exposure controls hold the prompt bank fixed. That distinction permits direct diagnosis without attributing between-method differences to one architectural cause.

Expressiveness and finite-bank selection ceilings. Visual prompt tuning and theoretical analyses study what prompting can express (Jia et al., 2022; Petrov et al., 2024; Le et al., 2026). Our substitutions test available trained prompts under specified readouts, rather than the capacity of all possible prompts. Label-informed oracle selection also has a history in dynamic classifier selection (Ko et al., 2008; Cruz et al., 2015). The exhaustive oracle used here follows that diagnostic role: it describes correct alternatives in a fixed bank, while a deployable policy must identify them without test labels (Shihab et al., 2026). Consequently, the ceiling is interpreted separately from domain-specialisation claims and from the evidence that classifier exposure changes component effects.