RETHINKING ZERO-SHOT EVALUATION IN CONTINUAL VISION-LANGUAGE PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual pretraining of vision-language models is commonly evaluated by zero-shot accuracy, which informs whether to retrain from scratch every year at about four times the compute or to keep all past data for replay. Accuracy drops during training are read as forgetting. Yet zero-shot accuracy conflates the image features with their alignment to class-name embeddings, while also depending on evaluator-chosen class names and templates. We disentangle representation quality from zero-shot alignment on all 28 released TiC-CLIP checkpoints by comparing zero-shot classification with nearest class mean and linear probing on the same frozen image features, while varying only the prompt. The checkpoints cover three training strategies: retraining from scratch on all data (*Oracle*), replaying all past data while continuing from the previous checkpoint (*Cumulative*), and sequential updating without replay (*Sequential*), across 2016–2022 and two data-filtering regimes. We find that zero-shot comparisons between training strategies are unstable, imprecise, and inflated, and that zero-shot declines alone do not establish representation forgetting. With model weights fixed, switching between two commonly used templates reverses the sign of the *Oracle–Cumulative* gap on 6 of 11 datasets under basic filtering. Across datasets, paired tests fail to distinguish *Oracle* from *Cumulative* using zero-shot accuracy under either data filter, but separate them using linear probes under both. Under the probe, replay is worth 0.6 to 1.5 points, a third to a fifth of the 3.2 to 4.4 that zero-shot accuracy reports, and its value grows at most a third as fast over time; retraining from scratch adds a further 0.6 to 0.7 points. Along the *Cumulative* trajectory, where the training pool only grows, zero-shot accuracy drops by at least one point on 10 of 132 one-year steps, and the linear probe on none of them. These findings motivate three checks for comparing continual-pretraining strategies from frozen checkpoints: probe the image features alongside the zero-shot head, test paired per-dataset differences rather than suite-level averages, and verify conclusions across standard prompt configurations. Without these checks, differences attributed to training may instead reflect the evaluation protocol.

1 INTRODUCTION

Contrastive vision-language models learn a shared embedding space for images and text, enabling zero-shot classification by comparing image embeddings with text embeddings of class names (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). As web data and deployment distributions evolve, continual pretraining offers a way to update these models without repeatedly training from scratch. TiC-CLIP (Garg et al., 2023) studies this setting using yearly data slices from 2016 to 2022. It compares continuing from the previous checkpoint with replay of all earlier data (*Cumulative*), continuing on the new year alone (*Sequential*), and retraining from scratch on all available data (*Oracle*). Its reported zero-shot results favor Cumulative as an efficient alternative to Oracle and show a benefit from replay over Sequential. Such comparisons inform consequential choices about training compute and the retention and reuse of historical data. Their interpretation therefore matters: what do zero-shot gaps between training strategies, and zero-shot declines during training, tell us about the capabilities these models retain?

Zero-shot accuracy measures the performance of a particular image-text decision rule. It depends jointly on the visual representation, its compatibility with text embeddings, and the class names and

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.

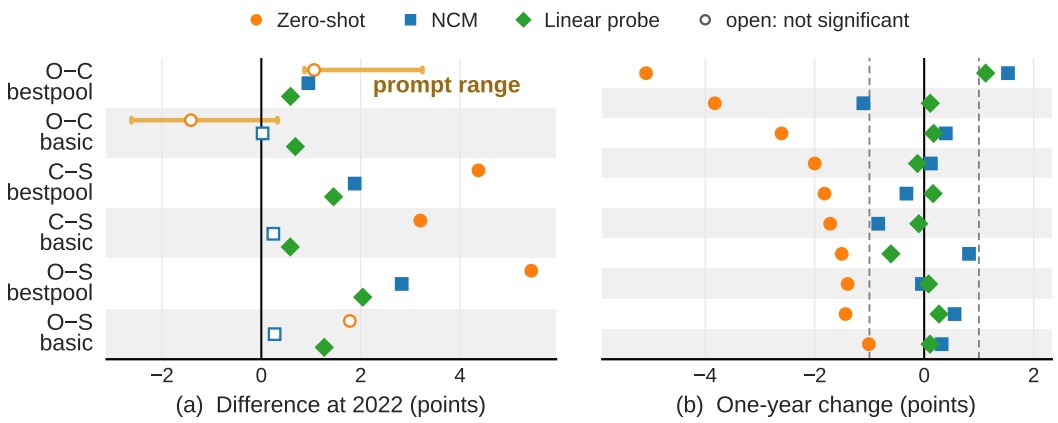


Figure 1: **Zero-shot comparisons are prompt-sensitive, and zero-shot declines need not imply comparable declines in visual discriminability.** (a) Final-checkpoint accuracy differences (Oracle minus Cumulative), averaged over eleven datasets. Open markers indicate differences that are not statistically significant ($p \geq 0.05$). The bar shows the range of mean zero-shot differences across five common class-name and template configurations (Table 11); this range crosses zero under basic filtering. (b) Accuracy changes for the 10 of 132 dataset-year transitions on which zero-shot accuracy drops by at least one percentage point, alongside NCM and linear-probe changes measured on the same frozen checkpoints. Dashed lines mark changes of ± 1 percentage point.

templates chosen by the evaluator. For a fixed task and prompt, a decline is a genuine loss of zero-shot performance. It does not, however, establish a corresponding loss of visual discriminability: features may remain useful for classification with an image-based readout even when their scores against class-name embeddings become less reliable. This distinction is relevant both to the interpretation of forgetting (McCloskey & Cohen, 1989; Goodfellow et al., 2013) and to downstream use. Visual encoders are often reused with classifiers trained on labelled images or incorporated into multimodal language models without their original text encoders (Alayrac et al., 2022; Liu et al., 2023; Chen et al., 2024). Conversely, preserving visual discriminability does not guarantee that a model retains its zero-shot capabilities. These are complementary properties, and an evaluation intended to guide continual pretraining should distinguish them.

We investigate this distinction using all 28 released TiC-CLIP checkpoints (Garg et al., 2023), evaluated on eleven classification datasets used in the prompt-learning literature (Zhou et al., 2022b;a) under both benchmark data filters. We apply two complementary controls. First, we hold model weights fixed and vary only the class names and prompt templates, testing whether comparisons between training strategies depend on the evaluation configuration. Second, we evaluate the same frozen visual representations using zero-shot classification, nearest class mean (NCM), and a linear probe. NCM and linear probing construct their classifiers from labelled images without using the text encoder. They therefore measure visual discriminability under two different readouts: one based on class means and the other on a learned linear decision boundary. These protocols do not provide a causal decomposition of representation and alignment changes. Instead, they test whether conclusions drawn from zero-shot evaluation are also supported by image-only evaluation.

Our first finding is that prompt choice can reverse comparisons between training strategies without any change to the models. Under basic filtering, switching from a generic template to the CLIP prompt ensemble reverses the sign of the Oracle versus Cumulative gap on 6 of the 11 datasets (Figure 1(a)). Across five common prompt configurations, this gap varies by a median of 3.9 percentage points, compared with a median gap magnitude of 1.8 points. Thus, the apparent advantage of one training strategy over another can depend substantially on an evaluator-controlled choice. A comparison made under one prompt configuration need not predict the ranking under another, even on the same dataset.

Our second finding is that zero-shot declines need not be accompanied by comparable declines in image-only classification. Along the Cumulative trajectory, with the prompt fixed, zero-shot accuracy drops by at least one percentage point on 10 of the 132 dataset-year transitions spanning eleven datasets, six annual transitions, and two filters. Four of these drops exceed or equal two points. Among the same ten transitions, NCM drops by at least one point only once, while the linear

probe never does (Figure 1(b)). These discrepancies occur even though Cumulative retains earlier data in its training pool. They do not show that the visual representation is unchanged or that no information has been lost. They show that a decline in zero-shot performance alone is insufficient to establish a decline in visual discriminability as measured by these two protocols.

Our third finding is that the magnitude and consistency of training-strategy gains depend strongly on the evaluation protocol. Linear probing supports the ordering Oracle above Cumulative above Sequential under both filters. In paired comparisons across datasets, the Oracle versus Cumulative difference is statistically significant under linear probing for both filters, whereas it is significant under neither filter for zero-shot classification and changes sign between them. Across datasets, zero-shot gaps also vary two to seven times more than their linear-probe counterparts. The gains measured by linear probing are modest: Oracle exceeds Cumulative by 0.6 to 0.7 percentage points, and replay improves over Sequential by 0.6 to 1.5 points. The latter is only a third to a fifth of the 3.2 to 4.4 points measured by zero-shot classification. These differences do not invalidate the zero-shot benefit of replay; they show that its magnitude cannot be directly interpreted as a gain in visual discriminability.

Together, these results support a broader evaluation of continual vision-language pretraining. On TiC-CLIP, image-only evaluation supports the reported ordering of training strategies while revealing smaller gains and a different pattern of performance changes over time. We recommend reporting NCM or linear-probe results alongside zero-shot accuracy, documenting the class names and templates used, and testing the sensitivity of strategy comparisons to common prompt configurations. Paired per-dataset differences and their uncertainty should accompany aggregate results. These checks require frozen checkpoints and labelled evaluation data, without additional pretraining. They help distinguish the benefits of a training strategy for zero-shot use from its benefits for visual representation reuse, providing a clearer basis for decisions about retraining and replay.

2 Related Work

2.1 Continual Pretraining of Vision-Language Models

Continual pretraining of CLIP-style models has been studied under both unconstrained web-scale time streams and more controlled or domain-oriented regimes. TiC-CLIP (Garg et al., 2023) established the dominant benchmark by partitioning DataComp (Gadre et al., 2023) into yearly slices and publicly releasing checkpoints for three training strategies. Its performance matrix is summarized by in-domain accuracy, backward and forward transfer, with the key finding that cumulative training with full replay recovers oracle accuracy to within roughly one point at substantially lower compute. The same protocol was extended to language models in TiC-LM (Li et al., 2025). Complementary settings include FoMo-in-Flux (Roth et al., 2024) and TIME (Dziadzio et al., 2025), which examine long-horizon adaptation under fixed FLOP budgets on a knowledge-accumulation versus zero-shot-retention plane; C-CLIP (Liu et al., 2025) tracks domain-incremental retrieval together with open-vocabulary classification; and earlier methods such as IncCLIP (Yan et al., 2022), CTP (Zhu et al., 2023). Domain-adaptive single- or few-stage variants (Liu et al., 2024; Stevens et al., 2024; Yao et al., 2025) follow the same template on specialized corpora.

2.2 Evaluation Protocols

Across this literature the reported metric is zero-shot classification or retrieval accuracy, and both are computed from class-name or caption embeddings produced by the frozen text encoder. Declines along training are measured the same way: FoMo-in-Flux (Roth et al., 2024) reports average zero-shot accuracy over a fixed set of held-out datasets at each step of the stream, and TiC-CLIP reports this alongside backward transfer on its year-wise retrieval tasks. In continual fine-tuning of CLIP on a sequence of classification datasets, largely those we use, the drop in zero-shot accuracy on datasets not yet trained on is attributed to catastrophic forgetting, and methods are designed to prevent it (Zheng et al., 2023; Yu et al., 2024). Linear probes (Alain & Bengio, 2016), nearest-class-mean classifiers (Mensink et al., 2013) and representation-similarity measures (Kornblith et al., 2019) appear only sporadically; Mod-X (Ni et al., 2023) is the principal exception that quantifies geometric disorder, yet still evaluates downstream performance through retrieval.

Table 1: (a) The three training strategies; t is the time step, one per year from 2016 to 2022. A warm start continues from the previous step's weights. (b) Downstream protocols ordered by how much of the model they are allowed to use or change. One protocol is named per level.

(a) Training strategies.

Method	Init.	Data at t
Sequential	warm	year t
Cumulative	warm	all $\leq t$
Oracle	random	all $\leq t$

(b) Downstream protocols.

Level	Protocol	Text	Features	Params
L0	ZS classification	used	frozen	none
L1	NCM	unused	frozen	none
L2	Linear probe	unused	frozen / tuned	yes

A drop in accuracy cannot faithfully reflect the changes in image features. Supervised class-incremental learning (CIL) shows this directly: the classifier drifts toward recently seen classes (Wu et al., 2019; Hou et al., 2019; Zhao et al., 2020), and probing the frozen features finds them far less damaged than the accuracy suggests (Davari et al., 2022). Continual learning of language models shows the same pattern: performance drops often reflect a loss of task alignment rather than of the underlying knowledge (Zheng et al., 2025). In CIL the classifier is a trainable head that can be re-balanced or probed independently of the representation. Under zero-shot evaluation it is not: the classifier is assembled at test time from class names, prompt templates, and the text encoder. A lower zero-shot accuracy may therefore come from degradation of the image features or from drift in their alignment with class-name embeddings, and the two are repaired in different places. Standard benchmarks report one accuracy and leave them entangled; separating them is what we measure.

3 SETUP

Evaluated models. We evaluate all 28 checkpoints released by the TiC-CLIP authors for TiC-DataComp (xlarge), all ViT-B/16. The benchmark partitions a web-scale image-text pool into yearly slices from 2016 to 2022 and defines the three training strategies summarized in Table 1a. Each of its two filtering regimes, *bestpool* and *basic*, provides 14 checkpoints: seven for *Cumulative* (2016–2022), six for *Sequential* (2017–2022), and one for *Oracle* (2022). Sequential shares the 2016 checkpoint with Cumulative because no earlier data are available for replay at that stage. For Cumulative and Sequential, we refer to the sequence of checkpoints from 2016 to 2022, including this shared starting point, as a trajectory. Oracle is available only at the final year.

The two filtering regimes differ in how they select training pairs. Basic filtering uses language, caption-length, and image-size criteria, while bestpool additionally uses scores from off-the-shelf CLIP models to select high-scoring pairs (Gadre et al., 2023). Sequential and Cumulative use the same compute budget but sample from different training pools: Sequential uses only the current year's data, whereas Cumulative includes all preceding years. The 2022 Oracle model is trained from scratch on all data available through 2022, using approximately four times the total compute. We report results under both filtering regimes. Per-checkpoint, per-dataset accuracies for all three evaluation protocols are provided in Tables 6–8 in the appendix.

Downstream protocols and datasets. Table 1b organizes evaluation protocols by the model components they use and the adaptation they allow. Our main distinction is between L0, which classifies images using text embeddings of class names, and L1/L2, which use the image features alone; within the latter, L1 fits no parameters (NCM (Mensink et al., 2013)), L2 learns a classifier on frozen features, as in linear probing (Alain & Bengio, 2016). Our classification experiments instantiate these levels with zero-shot classification, NCM, and linear probing, respectively.

We evaluate all three classification protocols on the same eleven datasets (Appendix Table 5), following the suite commonly used in the prompt-learning literature (Zhou et al., 2022b;a; Gao & Dong, 2025; Long et al., 2025). This choice enables matched comparisons: NCM and linear probing require labelled training examples, and all three protocols operate on a fixed set of classes within each task. Retrieval tasks do not fit this classification setting, while benchmark tasks without an official training split do not support the same training-based evaluation procedure. We selected the suite before running the evaluations and retained it throughout. All combinations of protocol, checkpoint, and filtering regime are evaluated on all eleven datasets, with no datasets omitted from aggregate results. We report per-dataset accuracies alongside aggregate results. Our conclusions therefore concern classification on this suite, rather than the full range of tasks in the original benchmark.

Table 2: **The same two checkpoints, compared under five prompt configurations.** Oracle minus Cumulative at 2022 in zero-shot accuracy (points), bestpool / basic. The weights are identical in every row. Local names are those shipped with each dataset; CLIP names and the CLIP template are those of (Radford et al., 2021); Generic is the single template `a photo of a {}`; Ensemble averages the full CLIP template set of each dataset. The last row is the configuration used in the rest of the paper. Range is the largest minus the smallest of the five differences on a dataset. Per-dataset values are in Table 11.

Names	Template	Mean Δ	Oracle leads (of 11)	ImageNet Δ
Local	Generic	$+3.2$ / -2.6	8 / 2	$+1.2$ / -2.4
CLIP	Generic	$+1.9$ / -1.7	8 / 3	$+1.2$ / -2.4
Local	CLIP	$+0.9$ / $+0.3$	7 / 7	$+0.9$ / -1.0
CLIP	CLIP	$+0.9$ / -2.2	7 / 5	$+0.9$ / -1.0
CLIP	Ensemble	$+1.1$ / -1.4	7 / 7	$+1.0$ / $+1.5$
Datasets whose sign changes across the five configurations: 1 / 6 of 11				
Median range across the five configurations: 0.7 / 3.9				
Median $\lvert\Delta\rvert$ under the CLIP ensemble: 1.0 / 1.8				

4 What Zero-Shot Comparisons Measure

Zero-shot accuracy is read in two ways during continual pretraining: a gap between checkpoints ranks training strategies, and a decline along training is taken as forgetting. We test each reading with one of the two controls introduced in Section 1. Section 4.1 holds the weights fixed and changes only the class names and template, and finds that this choice alone can reverse the ranking of two checkpoints. Section 4.2 holds the prompt fixed and follows the Cumulative trajectory, and finds zero-shot declines on steps where nearest class mean and the linear probe do not decline.

4.1 Prompt Sensitivity

Zero-shot accuracy is reported as a property of a checkpoint, but it also depends on the class names and templates, which the evaluator chooses. For a single model this dependence is well documented: class descriptions written by a language model raise accuracy (Menon & Vondrick, 2022), and random words in their place give comparable gains (Roth et al., 2023). This choice alone can decide which of two checkpoints ranks higher. In Table 2, we evaluate the final Oracle and Cumulative checkpoints under five configurations, with the weights fixed.[1] Under basic filtering, the configuration flips which model is ahead on 6 of the 11 datasets, and it flips the average as well, from Cumulative ahead by 2.6 points to Oracle ahead by 0.3. Under bestpool the ranking is more stable, but the average gap still more than triples, from 0.9 to 3.2 points. A ranking of checkpoints by zero-shot accuracy therefore holds only under the prompt configuration it was measured with.

These reversals are not driven by a few datasets with large prompt effects. On most datasets the prompt choice moves accuracy only a little: for a single checkpoint the median spread across the five configurations is 1.6 points, and the two name sets give identical accuracy on five of the eleven datasets. But the difference between Oracle and Cumulative is just as small. Under basic, its median size is 1.8 points while the prompt choice moves it by a median of 3.9, so it flips on six datasets, including ImageNet, where Cumulative leads by 2.4 points under the generic template and Oracle leads by 1.5 under the CLIP ensemble. Under bestpool the difference is larger than the prompt effect, with medians of 1.0 and 0.7, and it flips on only one dataset. A prompt effect of one or two points is small for accuracy but large for a comparison, because the gaps between checkpoints are of the same size.

4.2 Zero-Shot Declines Without Comparable Probe Declines

With the prompt fixed, zero-shot accuracy still falls on steps where the image features do not. We show this on the Cumulative trajectory, where each step adds a year of data to a pool from which

[1] All other results in this paper use the CLIP ensemble, the last row of Table 2: provided by Radford et al. (2021), which rewrite ambiguous labels and add a category word, with their full per-dataset template set averaged into one classifier. It is the configuration shipped with CLIP and, on average over the eleven datasets, the strongest of the five, so the zero-shot numbers we compare against are the most favourable of those available.

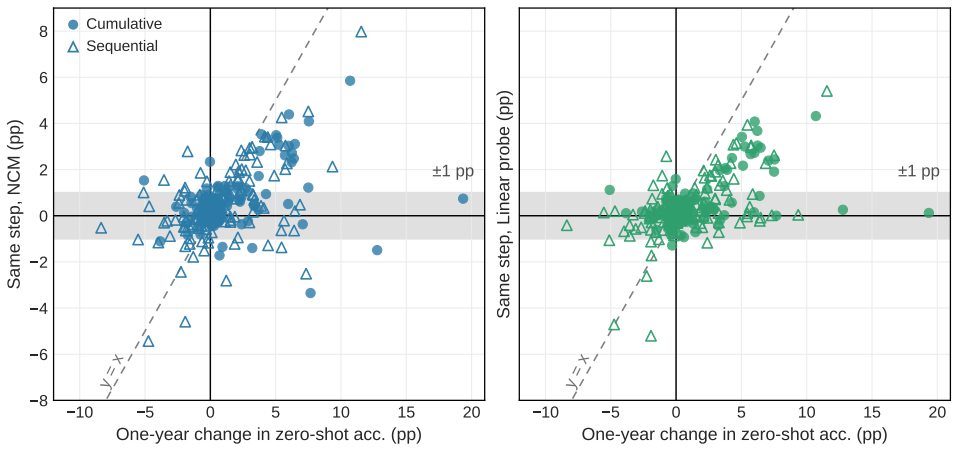


Figure 2: **Zero-shot declines need not coincide with comparable probe declines.** One-year accuracy changes on the same pair of frozen checkpoints, zero-shot against NCM (left) or linear probe (right), for all eleven datasets, six transitions and both filters. Filled markers denote Cumulative transitions, on which the transition statistics in Section 4.2 are based; open markers denote Sequential transitions, shown for reference. The dashed line $y = x$ marks equal accuracy changes under the two protocols. Points in the upper-left quadrant indicate a zero-shot decline accompanied by an increase in probe accuracy. Shaded bands mark changes within ± 1 percentage point of zero.

nothing is removed, so a drop along it cannot be blamed on data the model no longer sees. Training improves the image features and the alignment of class names with them together, so a rise in zero-shot accuracy does not say which of the two improved. A drop is more informative: if zero-shot accuracy drops on a step where nearest class mean and a linear probe on the same frozen weights do not, the loss lies in the alignment, not in the features.

Figure 1 (b) shows that zero-shot accuracy falls by at least one point on ten of the 132 one-year steps (11 datasets, 6 steps, 2 filters): nearest class mean falls as much on one of them and the linear probe on none, and on six nearest class mean rises instead. The pattern does not depend on the threshold. At half a point, 21 steps show a zero-shot decline and nearest class mean declines as much on 3 of them, the linear probe on 2; at two points, which exceeds the median width of the 95% confidence interval for a one-year zero-shot change (1.4 to 1.6 points, Table 14), four zero-shot drops remain and none under either image-side protocol (Table 13). Figure 2 shows all 132 steps; points in its upper-left quadrant, where zero-shot accuracy falls while nearest class mean rises, cannot arise from any monotone relation between the two protocols, and the six steps just described lie there.

When all earlier data is replayed, zero-shot accuracy falls on steps where the image-side protocols stay flat or even rise, so these drops come mainly from drift in image-text alignment rather than from degradation of the image features.[2]

5 Strategy Gaps under Image-Only Evaluation

The reported result (Garg et al., 2023) is that Oracle beats Cumulative, which beats Sequential. Under image-side protocols we measure the gaps between these strategies. This raises two key questions: which strategy is superior, and how large the gaps between them are. Section 5.1 shows that on the image features the ordering holds under both filters, while under zero-shot accuracy whether Oracle or Cumulative is ahead changes with the filter and the dataset. Section 5.2 shows that the gaps are also small on the image features: under the linear probe the largest is 2.0 points, and C−S is a third to a fifth of its zero-shot value.

[2]Neither image-side protocol is too coarse to see a loss. Over the same 132 steps each records declines of its own at one point, six under nearest class mean and one under the linear probe, but they fall on other steps: only one of the six coincides with a zero-shot decline and nor does that one. Nearest class mean also fits no parameters, so a loss in the features cannot be masked by refitting. The logit scale plays no part either: it does not change which class scores highest, and all 28 checkpoints share the same value (Appendix C).

We report differences between final checkpoints. O−C, Oracle minus Cumulative, is the cost of continual training relative to retraining from scratch. C−S, Cumulative minus Sequential, is the benefit of replay, and O−S combines the two. C−S is the cleanest comparison: both models are warm started, use the same compute, and differ only in the data they are trained on. Oracle is randomly initialized and retrained from scratch at every step, which over the seven steps costs about four times the compute of a continual run; any comparison involving it therefore mixes the training strategy with the initialization and the training schedule, and we flag this where it matters. Unless stated otherwise, values are means over the eleven datasets, reported as bestpool / basic. Differences are tested over the eleven datasets with the exact two-sided Wilcoxon signed-rank test.

5.1 Strategy Ordering and Statistical Significance

With a linear probe on the image features, the ordering Oracle > Cumulative > Sequential holds under both filters. Under zero-shot accuracy, only Cumulative over Sequential holds; whether Oracle or Cumulative is ahead changes with the dataset and the filter. The reason is that zero-shot differences are much less consistent across datasets: Table 3 shows that their standard deviation over the eleven datasets is two to seven times that under the linear probe. Per dataset the contrast is starker: in zero-shot accuracy the differences cover more than twenty points, while under the image-side protocols they stay within a few points of zero (Figure 4).

Oracle against Cumulative shows the difference. Under zero-shot accuracy, Oracle is ahead on seven datasets and behind on four under both filters, its average lead changes sign from $+1.1$ points under bestpool to -1.4 under basic, and the signed-rank test finds no consistent difference ($p = 0.21$ and 0.64). Under the linear probe, Oracle is ahead on eight and ten of the eleven datasets, is never behind by more than half a point, and the difference is significant under both filters with the same sign ($p = 0.01$). The sign of the gap flips between the two filters on four datasets under zero-shot accuracy and on one under the probe. Under basic, the Hodges-Lehmann interval for the zero-shot gap is $[-6.3, +1.9]$ points, eight points wide and containing zero, while for the probe it is $[+0.1, +1.2]$, one point wide and excluding it (Table 17). Both come from the same two checkpoints and the same eleven datasets. The same holds for the widest comparison on the suite: under basic, zero-shot accuracy does not establish that Oracle beats Sequential either ($p = 0.28$), while the linear probe does ($p = 0.004$).

Cumulative is ahead of Sequential on nearly every dataset, under zero-shot accuracy and on the image features alike. It leads on ten of the eleven datasets under zero-shot accuracy, and the difference is significant under both filters, under zero-shot accuracy ($p = 0.003$ and 0.002) and under the linear probe ($p = 0.001$ and 0.03). The one exception is nearest class mean under basic, where Cumulative leads by only 0.2 points while the standard deviation over the eleven datasets is 2.0, as large as under zero-shot accuracy; there the test finds no consistent difference ($p = 0.32$).

The ordering the benchmark reports is therefore right, but only the image features establish it for Oracle against Cumulative. The linear probe gives a consistent answer across the eleven datasets for gaps as small as 0.6 points, while zero-shot accuracy gives one for every comparison of 3.2 points or more and for none whose size is 1.8 points or less.

5.2 Magnitude of Strategy Gains

Table 3 shows that every gap between the three strategies is small on the image features. Under the linear probe, retraining from scratch is worth 0.6 / 0.7 points over continual training with replay, replay is worth 1.5 / 0.6 points over training on each year alone, and the two together separate the cheapest strategy from the most expensive by 2.0 / 1.3 points. Nearest class mean gives 0.0 to 0.9, 0.2 to 1.9 and 0.3 to 2.8. These are differences between accuracies of 82 to 87 points.

Zero-shot accuracy reports the same comparisons as larger. For C−S it gives $+4.4$ / $+3.2$ points, three to five times the $+1.5$ / $+0.6$ of the linear probe, and for O−S it gives $+5.4$ / $+1.8$ against $+2.0$ / $+1.3$. For O−C the difference is not one of size but of sign: $+1.1$ / -1.4 under zero-shot accuracy against $+0.6$ / $+0.7$ under the probe. Retraining from scratch brings less than a point over continual training measured on the image features, and it is trained with about four times the compute.

Learning the prompt instead of choosing it brings zero-shot accuracy close to the probe. With CoOp, Oracle minus Cumulative is $+1.0$ / $+0.8$ points and Cumulative minus Sequential $+2.2$ /

Table 3: **Three comparisons under three protocols.** Difference in accuracy between final checkpoints (points), bestpool / basic. O−C (Oracle minus Cumulative) is the cost of continual training relative to retraining from scratch, C−S (Cumulative minus Sequential) the benefit of replay, and O−S the two combined. NCM and LP use the full training split. *Mean* Δ and *SD* are the mean and standard deviation over the eleven datasets. *Leads* counts datasets where the first model is ahead. *Flips* counts datasets where the difference has opposite signs under the two filters. p is the exact two-sided Wilcoxon signed-rank test over the eleven datasets. Per-dataset values are in Table 17.

	Protocol	Mean Δ	SD	Leads (of 11)	Flips (of 11)	p
O−C	ZS	$+1.1$ / -1.4	2.07 / 6.12	7 / 7	4	0.21 / 0.64
	NCM-full	$+0.9$ / $+0.0$	0.86 / 1.19	10 / 8	2	0.002 / 0.64
	LP-full	$+0.6$ / $+0.7$	0.49 / 0.86	8 / 10	1	0.01 / 0.01
C−S	ZS	$+4.4$ / $+3.2$	2.76 / 2.07	10 / 10	0	0.003 / 0.002
	NCM-full	$+1.9$ / $+0.2$	1.05 / 2.00	10 / 9	3	0.002 / 0.32
	LP-full	$+1.5$ / $+0.6$	1.08 / 0.79	11 / 8	3	0.001 / 0.03
O−S	ZS	$+5.4$ / $+1.8$	3.51 / 7.28	10 / 9	1	0.003 / 0.278
	NCM-full	$+2.8$ / $+0.3$	1.79 / 2.64	11 / 8	3	0.001 / 0.206
	LP-full	$+2.0$ / $+1.3$	1.47 / 1.43	10 / 9	2	0.002 / 0.004

$+1.1$, against $+1.1$ / -1.4 and $+4.4$ / $+3.2$ under zero-shot accuracy (Appendix G). The image features are the same in both, so the extra zero-shot gap comes from the text side.

These averages depend on which datasets the suite contains, and much more so under zero-shot accuracy. Dropping any one of the 11 datasets moves the O−C average by at most 0.2 points under the LP and 0.3 under NCM, but by up to 1.4 under zero-shot accuracy (Table 18). The image-side averages are a property of the models; the zero-shot average is in part a property of the suite.

6 Replay Benefits over Time

Deciding whether to keep a pool that grows every year needs two further things: how that value changes as training continues, and how large it is next to what continual training gains over the same period. On the image features it grows at most a third as fast as zero-shot accuracy suggests, and it comes to between a tenth and a half of what the Cumulative trajectory gains from 2016 to 2022. We also ask whether replay brings robustness beyond accuracy, and find no consistent evidence that it does. Throughout we compare Cumulative against Sequential, the clean comparison of Section 5.

6.1 Temporal Dynamics

Under zero-shot accuracy the benefit of replay grows steadily. Figure 3 plots the median replay gap over datasets at each time step (values in Table 19). The median gap over the eleven datasets rises by 0.6 / 0.5 points per year, from 1.0 / 0.6 points at 2017 to 5.2 / 3.1 at 2022. On the image features it barely grows. Under bestpool the gap stays below one point for the first five steps and reaches 1.7 under the linear probe and 2.0 under nearest class mean only at 2022, giving slopes of 0.2 per year under both. Under basic it stays near half a point throughout, with slopes of 0.0 and 0.1.

The growth rate is what decides how long a pipeline without replay can run before it has to be reset, and ZS accuracy overstates it by a factor of three or more. Accuracy points are not strictly comparable across protocols that sit at different accuracy levels, so the next subsection removes that dependence, first by restating the gap in level-free units and then by matching the levels directly.

6.2 Relative Magnitude

On the image features, continual training gains little over the six years, and replay accounts for a large part of what it does gain. The median gain of the Cumulative trajectory from 2016 to 2022 is 4.9 / 4.6 points under nearest class mean and 3.2 / 2.5 under the linear probe. Taken dataset by dataset, the replay gap is a median 41 / 17 percent of that gain under nearest class mean and 47 / 9 under the linear probe, against 67 / 35 under zero-shot accuracy. The same gap can be read as a delay: on the image features, Sequential's 2022 model reaches where Cumulative had five years earlier. The unit saturates on a six-year run, so we report it only in Appendix 21.

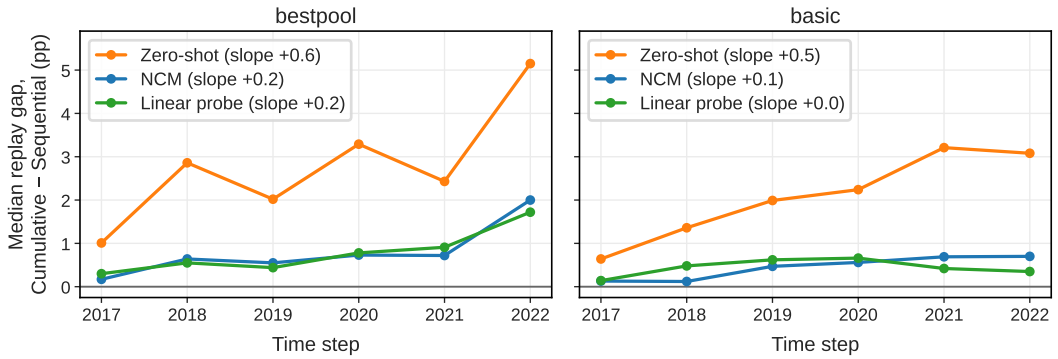


Figure 3: **The value of replay over time, under three protocols.** Median over the eleven datasets of Cumulative minus Sequential at each time step, in points. The two trajectories share the 2016 checkpoint, so the comparison starts at 2017. Slopes in the legend are the median over datasets of the per-dataset least-squares slope on time, in points per year.

Table 4: **The cost of dropping replay in level-free units.** *Norm. gap*: median of $(\mathrm{Cum}_{2022} - \mathrm{Seq}_{2022})/(\mathrm{Cum}_{2022} - \mathrm{Cum}_{2016})$ over the nine to eleven datasets with positive progress. *Matched*: the image-side protocol is interpolated to the zero-shot accuracy level on the n datasets where it can be, and $\Delta_{\mathrm{ZS}} - \Delta@k^\star$ is the median gap reduction there (exact Wilcoxon signed-rank; Table 20).

	Bestpool			Basic		
Protocol	Norm. gap	n	$\Delta_{\mathrm{ZS}} - \Delta@k^\star$	Norm. gap	n	$\Delta_{\mathrm{ZS}} - \Delta@k^\star$
ZS	0.67	—	—	0.35	—	—
NCM-full	0.41	7	$+3.24$ ($p = .031$)	0.17	6	$+2.30$ ($p = .031$)
LP-full	0.47	9	$+3.06$ ($p = .004$)	0.09	8	$+1.71$ ($p = .008$)

Matching the protocols at equal accuracy gives the same answer by a route that rescales nothing: the replay gap is smaller than the zero-shot gap on 29 of the 30 dataset-protocol pairs where the two can be matched, by a median of 1.7 to 3.2 points. The 14 pairs that cannot be matched are those where one shot already exceeds the zero-shot level, or where the full training split never reaches it.

6.3 Robustness Checks

Under the zero-shot head, replay appears to add robustness beyond accuracy: it raises ImageNet-C (Hendrycks & Dietterich, 2019) accuracy by 1.7 / 0.5 points more than ImageNet accuracy. Measured on a linear probe over the same features, and against the trend by which shifted accuracy rises with ImageNet accuracy, that excess survives only under basic filtering, and on natural shifts neither replay nor retraining from scratch has an advantage (Appendix F). We therefore do not claim a robustness benefit for replay beyond the accuracy it brings.

7 Conclusions and Limitations

Zero-shot accuracy depends jointly on visual representations, their alignment with text embeddings, and evaluator-selected prompts, so a gap between checkpoints does not by itself identify what changed in the model. On TiC-CLIP, we find that prompt choice can reverse strategy rankings and that zero-shot declines need not coincide with comparable declines under image-only probes. For the released checkpoints, linear probing supports the benchmark's strategy ordering under both filtering regimes, while measuring smaller gains and less variable strategy gaps across datasets. These findings motivate three complementary checks: evaluate frozen image features alongside zero-shot performance, report paired per-dataset differences and their uncertainty, and assess sensitivity to prompt configurations. Our findings concern the released TiC-CLIP checkpoints on eleven classification datasets; generalization across training seeds, model scales, and temporal retrieval tasks remains to be established. Together, our results provide a clearer basis for assessing retraining and replay by distinguishing their benefits for zero-shot performance from their benefits for visual discriminability.

AI USE STATEMENT

In this work, we used generative AI tools for interpreting results, implementing methods and assisting with translation. We did not use generative AI to help develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, to assist in the writing of proofs, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments. And generate synthetic data sets, clean and reformat dataset, support qualitative and thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools for summarizing existing literature and editing parts of the manuscript to improve readability. We have reviewed all AI-assisted work. All AI-assisted text, code and figures were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects, collects no new data, and releases no new model. It uses checkpoints and datasets that are already public, under their existing terms. The paper is a reanalysis of a published benchmark: what it questions is a measurement convention that the field has adopted, not the conduct of the benchmark's authors, whose released checkpoints made this study possible and whose central conclusion our image-side measurements support. We have tried to state the scope of our claims and the cases where our evidence is weaker in Section 7.

REPRODUCIBILITY STATEMENT

This paper trains no models. Every result is computed from the checkpoints released by the TiC-CLIP (Garg et al., 2023) for TiC-DataComp (xlarge), described in Section 3, and evaluated on the eleven public datasets listed in Table 5. Appendix A gives the implementation of each protocol: feature extraction and pre-processing, the nearest-class-mean prototypes, the linear probe together with its regularization grid and selection rule, the k-shot sampling and its seeds, the CoOp hyperparameters, and the significance tests. Table 24, Table 15 and Table 16 report the accuracy of every released checkpoint on every dataset under all three protocols, so every average, one-year change and slope in this paper can be re-derived from the appendix alone.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16691–16700. IEEE, 2022.

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

Sebastian Dziadzio, Vishaal Udandarao, Karsten Roth, Ameya Prabhu, Zeynep Akata, Samuel Albanie, and Matthias Bethge. How to merge your multimodal models over time? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20479–20491. IEEE, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Mengyu Gao and Qiulei Dong. Causality-guided prompt learning for vision-language models via visual granulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1141–1151. IEEE, 2025.

Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. Tic-clip: Continual training of clip models. *arXiv preprint arXiv:2310.16226*, 2023.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 831–839. IEEE, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, et al. Tic-lm: A web-scale benchmark for time-continual llm pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32231–32273, 2025.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, volume 2025, pp. 46461–46477, 2025.

Sifan Long, Zhen Zhao, Junkun Yuan, Zichang Tan, Jiangjiang Liu, Jingyuan Feng, Shengsheng Wang, and Jingdong Wang. Mutual prompt leaning for vision language models. *International Journal of Computer Vision*, 133(3):1258–1276, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

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

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013.

Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *International Conference on Machine Learning*, pp. 26129–26149. PMLR, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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

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

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

Karsten Roth, Vishaal Udandarao, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Matthias Bethge, and Zeynep Akata. A practitioner's guide to continual multimodal pretraining. *arXiv preprint arXiv:2408.14471*, 2024.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424. IEEE, 2024.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL `https://arxiv.org/abs/2007.00644`.

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

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 374–382. IEEE, 2019.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*, pp. 22–38. Springer, 2022.

Yuang Yao, Ruiqi Wu, Yi Zhou, and Tao Zhou. Continual retinal vision-language pre-training upon incremental imaging modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 111–121. Springer, 2025.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230. IEEE, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13205–13214. IEEE, 2020.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations*, volume 2025, pp. 67081–67146, 2025.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19068–19079. IEEE, 2023.

Table 5: The eleven datasets used for every classification result in this paper. The suite is the one standard in the prompt-learning literature and is dictated by what the image-side protocols require: a labelled training split and a single fixed label set per task. It was fixed before any evaluation was run and never changed. Every protocol, checkpoint and filtering regime is evaluated on all eleven datasets, and no dataset is dropped from any average.

Dataset	Classes	n_{train}	n_{test}
ImageNet (Deng et al., 2009)	1000	1,281,167	50,000
Caltech101 (Fei-Fei et al., 2004)	100	5,777	2,465
OxfordPets (Parkhi et al., 2012)	37	3,680	3,669
StanfordCars (Krause et al., 2013)	196	6,515	8,041
Flowers102 (Nilsback & Zisserman, 2008)	102	5,726	2,463
Food101 (Bossard et al., 2014)	101	75,750	25,250
FGVCAircraft (Maji et al., 2013)	100	6,667	3,333
SUN397 (Xiao et al., 2010)	397	19,850	19,850
DTD (Cimpoi et al., 2014)	47	3,760	1,880
EuroSAT (Helber et al., 2019)	10	18,900	8,100
UCF101 (Soomro et al., 2012)	101	9,537	3,783

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Hongguang Zhu, Yunchao Wei, Xiaodan Liang, Chunjie Zhang, and Yao Zhao. Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22200–22210. IEEE, 2023.

A Evaluation setup

This section fixes the evaluated datasets and the protocols. Everything in the rest of the appendix is measured on the eleven datasets of Table 5 under the protocols of Section A.1.

A.1 Implementation details

Features. All protocols extract the image features of the same frozen checkpoint. Features are taken from the projected image embedding (the pooled visual-tower output after the visual projection), with dimension $d = 512$. Images are preprocessed with the standard CLIP transform (resize to 224×224 with bicubic interpolation, centre crop, and normalization with the CLIP mean and standard deviation); all checkpoints share the ViT-B/16 architecture and use this same preprocessing. Features are ℓ_2-normalized before being passed to any protocol, and the same cached features are reused by nearest class mean, the linear probe and CoOp, so no difference between protocols can come from feature extraction.

Nearest class mean. Each class prototype is the mean feature of that class's training images, and a test image is assigned to the nearest prototype under cosine similarity. Prototypes are computed after normalization (the features entering the mean are already ℓ_2-normalized) and are re-normalized afterwards. NCM-full uses the entire training split of each dataset (for Caltech101, Flowers102, SUN397 and EuroSAT, the training segment from (Zhou et al., 2022b); on ImageNet, the full training set). Nearest class mean fits no parameters, so NCM-full is deterministic given the weights and the split; we therefore report no seed variation for it, and the across-seed standard deviation is exactly zero in the `full` rows of Tables 15 and 16.

Linear probe. LP-full fits multinomial logistic regression on the frozen features of the full training split (for Caltech101, Flowers102, SUN397 and EuroSAT, the training segment from (Zhou et al., 2022b)) with a GPU L-BFGS solver (torch, `max_iter` $= 200$, with strong-Wolfe line search in the full-data path). The regularization strength is selected on the protocol's validation split — the official validation segment where one exists, the validation segment from (Zhou et al., 2022b) for Caltech101, Flowers102, SUN397 and EuroSAT, and a deterministic fixed-seed 10% held-out slice of the training split otherwise — over a coarse logarithmic grid $\{10^6, 10^4, 10^2, 1, 10^{-2}, 10^{-4}, 10^{-6}\}$ followed by eight bisection steps on $[0.1C^*, 10C^*]$, and the probe is then refit on the full training split at the selected value. The probe has a bias term (unpenalized, as in `sklearn`) and the features are not standardized separately (they enter ℓ_2-normalized). The same grid and the same selection rule are used for every checkpoint and every dataset, so the probe cannot be tuned differently for different checkpoints.

k-shot protocols. For $k \in \{1, 4, 8, 16\}$ we sample k images per class from the training split. Within a seed, all six final checkpoints use the same k-shot subset, so the comparison between checkpoints is paired, and the same subsets are shared by the k-shot nearest class mean, linear probe and CoOp results. We use three seeds and report the mean together with the across-seed standard deviation averaged over the eleven datasets. With k shots the regularization strength of the probe is chosen on a per-class validation set following the CoOp protocol (one image per class for $k = 1$, four for $k \geq 4$), drawn from validation segment for the re-split datasets (Zhou et al., 2022b) and otherwise sampled from the training pool with the selected shots excluded under an independent seed; the value is used directly, without refitting.

CoOp. CoOp freezes both towers and learns only the context vectors prepended to each class name. The image features are therefore unchanged by construction. We use 16 context tokens, shared across all classes, initialized from the embedding of `a photo of a` (the remaining tokens start at zero), with the class name placed at the end. Training uses SGD with momentum 0.9, learning rate 2×10^{-3} with a cosine schedule and batch size 32 for the k-shot runs (512, or 4096 on ImageNet, for the full-data runs), for up to 200 epochs with no warmup (the official one-epoch constant warmup at 10^{-5} is omitted as negligible); early stopping selects the context vectors by validation accuracy, and the best-validation ones are restored for test. No augmentation is applied (all protocols read pre-extracted features). Full-data CoOp uses the same training split as LP-full; k-shot CoOp uses the shared subsets described above and is averaged over three seeds.

Significance tests. Differences over the eleven datasets are tested with the exact two-sided Wilcoxon signed-rank test, and interval estimates are exact Hodges-Lehmann intervals. Within a single dataset, the significance of a one-year change is tested with McNemar's exact test on the shared test set and with a paired bootstrap over test examples (10,000 resamples).

B PER-CHECKPOINT, PER-DATASET RESULTS

Table 6, Table 7 and Table 8 give the accuracy of every released checkpoint on every dataset under each of the three protocols. Every average, one-year change and slope in the paper is computed from the unrounded values behind these three tables and rounded once, so it may differ by 0.1 from the same quantity computed from the rounded entries.

C PROMPT CONFIGURATION

This section supports Section 4.1. Table 10 keeps the weights fixed and varies only the class names and the template. Table 11 does the same for a comparison between two checkpoints. The learned temperature as one alternative explanation is ruled out directly: the logit scale is bitwise identical across all 28 checkpoints (the ln100 cap CLIP clamps to), so it cannot distinguish any checkpoint from any other.

Table 6: Zero-shot accuracy (%) of every released checkpoint along the TiC-DataComp (XL) trajectory. *Shared*: the first-step 2016 checkpoint from which both Cumulative and Sequential continue; Oracle exists only for 2022. Mean and Med. are over the eleven datasets. IN: ImageNet, Cal: Caltech101, Pet: OxfordPets, Car: StanfordCars, Flo: Flowers102, Food: Food101, Air: FGVCAircraft, SUN: SUN397, ESAT: EuroSAT, UCF: UCF101.

Method	Year	IN	Cal	Pet	Car	Flo	Food	Air	SUN	DTD	ESAT	UCF	Mean	Med.
Bestpool														
Shared	2016	62.9	86.4	86.3	73.9	69.4	84.1	15.9	60.3	46.4	35.3	56.1	61.6	62.9
Cum.	2017	68.0	86.5	90.6	81.4	73.6	87.9	21.9	66.8	53.5	33.6	61.8	66.0	68.0
Cum.	2018	69.1	86.3	90.2	83.8	73.0	88.0	22.3	68.2	55.6	53.0	62.6	68.4	69.1
Cum.	2019	69.9	87.0	89.9	86.5	72.6	88.5	22.3	67.9	58.0	53.2	63.7	69.0	69.9
Cum.	2020	70.3	86.5	90.2	87.1	73.4	88.5	25.2	68.4	58.7	53.8	63.4	69.6	70.3
Cum.	2021	70.6	86.5	90.6	87.1	73.5	88.8	27.3	67.8	58.0	57.5	62.8	70.1	70.6
Cum.	2022	70.8	86.5	90.0	87.7	72.1	88.7	26.8	67.9	58.2	53.6	64.1	69.7	70.8
Seq.	2017	67.5	85.9	89.5	81.4	72.6	86.7	20.8	66.1	50.3	41.7	58.2	65.5	67.5
Seq.	2018	66.8	85.9	86.0	83.6	70.6	85.1	19.2	66.8	51.9	47.2	59.7	65.7	66.8
Seq.	2019	67.2	86.0	89.3	86.6	72.5	85.9	23.4	65.9	51.6	48.0	60.5	67.0	67.2
Seq.	2020	67.0	85.8	88.6	86.7	70.1	85.9	21.7	65.7	53.5	47.5	58.6	66.5	67.0
Seq.	2021	67.1	86.1	89.1	84.8	71.6	86.3	23.9	65.0	52.1	49.9	61.6	67.1	67.1
Seq.	2022	65.6	85.1	83.6	80.8	69.6	84.4	19.1	64.1	52.0	55.4	59.0	65.3	65.6
Ora.	2022	71.8	86.5	92.8	88.1	77.4	89.9	25.4	67.0	58.8	52.9	67.6	70.7	71.8
Basic														
Shared	2016	53.2	81.5	76.5	64.9	58.5	79.0	10.4	57.9	42.0	26.3	47.1	54.3	57.9
Cum.	2017	59.2	82.8	82.8	75.6	64.9	84.1	11.2	63.2	49.5	33.9	53.3	60.0	63.2
Cum.	2018	58.9	82.9	82.6	79.5	62.9	84.0	11.3	58.1	48.5	32.5	51.8	59.4	58.9
Cum.	2019	61.1	83.5	81.9	83.2	65.2	84.5	12.3	64.1	51.8	35.7	55.2	61.7	64.1
Cum.	2020	61.8	83.7	83.2	84.6	65.8	84.3	12.9	64.6	52.0	36.4	55.0	62.2	64.6
Cum.	2021	62.1	84.2	82.5	85.0	64.0	84.4	12.1	65.1	51.7	49.2	54.3	63.1	64.0
Cum.	2022	62.3	83.9	85.4	85.2	64.8	84.1	11.8	65.7	51.1	46.6	55.8	63.3	64.8
Seq.	2017	58.6	83.6	82.3	76.4	64.4	82.0	8.5	62.3	48.8	27.5	50.1	58.6	62.3
Seq.	2018	57.6	83.3	78.7	80.0	56.0	80.7	7.8	60.4	43.7	36.8	51.7	57.9	57.6
Seq.	2019	58.5	83.5	79.9	82.7	62.5	80.8	9.4	64.5	48.1	44.2	53.7	60.7	62.5
Seq.	2020	58.5	83.1	82.3	82.4	57.8	80.3	11.3	63.9	47.9	42.2	51.9	60.1	58.5
Seq.	2021	58.9	83.8	81.2	82.2	59.9	79.5	10.6	63.9	44.6	41.3	50.5	59.7	59.9
Seq.	2022	59.1	81.9	83.6	79.1	57.5	80.6	8.3	62.7	48.0	46.9	53.7	60.1	59.1
Ora.	2022	63.8	84.1	85.0	83.8	66.6	85.2	13.7	54.1	53.5	31.5	59.8	61.9	63.8

Table 7: Nearest-class-mean accuracy (%) along the trajectory (image-derived prototypes; the text encoder is not used). Abbreviations as in Table 6.

Method	Year	IN	Cal	Pet	Car	Flo	Food	Air	SUN	DTD	ESAT	UCF	Mean	Med.
Bestpool														
Shared	2016	63.4	94.7	81.5	82.4	96.2	83.7	46.5	68.3	72.7	85.2	74.4	77.2	81.5
Cum.	2017	66.9	95.3	84.9	86.5	97.1	86.5	49.6	71.5	72.3	84.4	77.0	79.3	84.4
Cum.	2018	67.9	95.2	85.7	87.3	97.4	86.9	50.1	72.3	73.5	85.1	77.1	79.9	85.1
Cum.	2019	68.4	96.1	85.6	88.5	97.6	87.0	52.4	72.7	73.1	84.7	78.4	80.4	84.7
Cum.	2020	69.0	96.0	85.4	88.8	97.6	87.3	52.2	73.0	72.9	85.9	78.4	80.6	85.4
Cum.	2021	69.4	96.5	86.1	88.8	97.7	87.7	53.5	73.4	73.1	86.2	79.0	81.0	86.1
Cum.	2022	69.7	96.0	86.4	89.5	97.7	87.8	54.2	73.4	73.4	85.1	79.4	81.1	85.1
Seq.	2017	66.5	95.1	84.5	86.9	97.1	86.4	48.3	71.4	73.1	84.6	76.0	79.1	84.5
Seq.	2018	66.7	95.6	84.2	87.9	97.1	85.1	47.8	71.9	72.9	83.9	77.5	79.1	83.9
Seq.	2019	67.4	95.6	84.3	89.4	97.0	85.8	51.2	72.2	74.4	84.8	77.8	80.0	84.3
Seq.	2020	68.1	96.1	84.1	88.7	97.6	85.9	54.0	72.3	73.1	83.3	77.2	80.0	83.3
Seq.	2021	68.7	96.3	85.3	88.5	97.6	86.5	55.9	71.9	72.8	85.3	76.9	80.5	85.3
Seq.	2022	67.7	96.1	84.3	87.3	97.1	85.2	50.5	71.5	71.6	84.0	76.7	79.3	84.0
Ora.	2022	71.0	96.3	87.6	89.8	97.8	89.0	57.3	74.3	73.4	86.0	80.6	82.1	86.0
Basic														
Shared	2016	54.3	90.3	63.8	73.2	92.7	78.8	40.4	64.7	67.3	81.4	68.4	70.5	68.4
Cum.	2017	58.7	91.4	66.2	79.0	95.1	82.2	40.3	67.8	68.5	78.1	71.1	72.6	71.1
Cum.	2018	59.9	91.9	66.6	82.6	95.3	82.9	39.9	69.3	68.8	78.6	71.9	73.4	71.9
Cum.	2019	60.5	92.3	67.5	84.3	95.9	82.7	38.6	69.8	69.3	77.2	72.4	73.7	72.4
Cum.	2020	61.1	92.6	68.1	85.0	95.6	83.3	39.8	69.7	69.5	75.5	73.3	74.0	73.3
Cum.	2021	61.5	93.1	68.0	85.4	95.3	83.4	39.0	70.1	69.4	74.0	72.9	73.8	72.9
Cum.	2022	62.2	93.0	69.0	85.0	95.3	83.5	38.8	70.5	69.3	74.4	74.2	74.1	74.2
Seq.	2017	58.6	92.0	65.9	81.2	94.9	81.8	35.9	68.1	67.8	78.6	71.0	72.3	71.0
Seq.	2018	59.3	92.5	67.4	83.5	94.4	80.0	34.7	69.2	68.8	80.7	70.8	72.9	70.8
Seq.	2019	60.1	92.4	67.7	85.5	94.6	80.3	36.6	69.5	67.5	78.2	71.6	73.1	71.6
Seq.	2020	60.8	92.0	68.8	84.9	95.0	80.5	38.9	69.4	68.1	79.4	72.3	73.6	72.3
Seq.	2021	61.2	92.4	68.9	83.9	94.1	81.4	40.7	69.7	67.9	78.9	71.8	73.7	71.8
Seq.	2022	61.9	91.9	71.8	83.0	95.0	81.2	38.3	69.6	68.6	78.7	72.7	73.9	72.7
Ora.	2022	62.9	93.3	70.3	85.1	95.6	83.6	40.4	71.2	68.0	71.8	73.3	74.1	71.8

Table 8: Full-data linear-probe accuracy (%) along the trajectory on frozen image features. Abbreviations as in Table 6.

Method	Year	IN	Cal	Pet	Car	Flo	Food	Air	SUN	DTD	ESAT	UCF	Mean	Med.
Bestpool														
Shared	2016	65.9	96.1	89.0	88.0	98.7	89.4	61.4	73.0	77.8	96.6	80.7	83.3	88.0
Cum.	2017	69.4	96.8	91.5	90.4	99.3	91.5	64.1	76.0	80.2	96.5	83.4	85.4	90.4
Cum.	2018	70.0	96.4	91.0	91.6	99.2	91.7	63.3	76.7	80.5	96.6	84.1	85.5	91.0
Cum.	2019	70.5	96.9	91.7	92.0	99.1	91.7	64.8	76.9	80.7	96.9	84.3	86.0	91.7
Cum.	2020	70.8	97.0	92.0	92.2	99.2	92.1	65.1	77.4	81.7	96.7	84.5	86.3	92.0
Cum.	2021	71.4	97.4	92.3	92.4	99.3	92.3	65.0	77.5	81.3	96.7	85.4	86.5	92.3
Cum.	2022	72.1	97.2	91.7	92.3	99.4	92.4	65.1	77.6	81.0	96.8	85.2	86.4	91.7
Seq.	2017	69.1	97.1	91.2	90.6	98.8	91.2	61.1	76.1	78.9	96.4	82.7	84.8	90.6
Seq.	2018	69.3	96.8	90.4	91.6	98.8	90.5	61.1	76.1	79.8	96.4	83.9	85.0	90.4
Seq.	2019	69.5	96.5	89.6	92.3	99.1	91.0	63.4	76.4	80.2	97.0	84.6	85.4	89.6
Seq.	2020	70.1	97.0	90.8	92.0	99.2	91.1	64.7	76.5	79.9	96.6	82.9	85.5	90.8
Seq.	2021	70.5	97.0	91.0	91.7	99.2	91.2	66.4	76.5	80.1	96.5	83.6	85.8	91.0
Seq.	2022	69.8	97.2	91.2	91.1	99.0	90.1	61.7	75.9	79.2	96.7	83.1	85.0	90.1
Ora.	2022	72.8	97.2	92.8	92.8	99.4	93.0	66.5	78.6	81.9	96.8	85.6	87.0	92.8
Basic														
Shared	2016	57.1	94.3	83.1	82.6	97.5	87.5	54.0	70.9	75.3	96.1	78.5	79.7	82.6
Cum.	2017	61.1	95.3	86.7	86.9	98.4	89.6	53.8	73.9	77.2	96.1	81.6	81.9	86.7
Cum.	2018	62.4	94.9	86.1	88.9	98.3	90.0	54.4	75.0	77.3	96.4	80.9	82.2	86.1
Cum.	2019	63.0	95.1	86.7	90.1	98.4	90.0	54.4	75.2	79.3	95.9	80.8	82.6	86.7
Cum.	2020	63.4	95.2	86.4	90.8	98.2	89.9	53.4	75.3	79.7	95.8	81.3	82.7	86.4
Cum.	2021	64.0	95.5	87.0	90.9	98.3	90.2	54.2	75.8	78.9	96.0	81.8	83.0	87.0
Cum.	2022	64.3	95.5	87.2	91.0	98.3	89.9	52.9	75.6	79.0	96.2	81.1	82.8	87.2
Seq.	2017	61.0	94.9	86.0	88.0	98.5	89.4	48.8	74.0	77.6	96.1	80.9	81.4	86.0
Seq.	2018	61.9	95.5	85.6	89.5	98.1	88.4	49.7	74.2	76.5	96.1	80.4	81.5	85.6
Seq.	2019	62.4	94.7	86.0	91.2	98.0	88.7	50.2	74.9	78.1	96.2	79.9	81.9	86.0
Seq.	2020	63.0	94.5	84.8	90.7	98.2	88.7	50.6	74.7	77.3	96.1	80.7	81.8	84.8
Seq.	2021	63.6	95.2	86.6	90.5	98.2	88.9	53.1	75.1	77.4	95.8	81.1	82.3	86.6
Seq.	2022	63.9	94.6	87.8	89.9	98.4	88.8	50.5	75.0	78.7	96.2	80.8	82.2	87.8
Ora.	2022	65.3	95.7	88.7	91.0	98.4	90.8	55.6	76.5	79.3	95.7	81.7	83.5	88.7

Table 9: **Per-dataset accuracy under four protocols** (%). ZS, NCM and LP are reproduced from Section B for reference. CoOp freezes both towers and re-learns only the text prompt, so the image features, and therefore the NCM and LP columns, are unchanged by construction.

		Bestpool				Basic			
Dataset	Method	ZS	NCM	LP	CoOp	ZS	NCM	LP	CoOp
ImageNet	Seq.	65.6	67.7	69.8	72.3	59.1	61.9	63.9	67.1
	Cum.	70.8	69.7	72.1	75.8	62.3	62.2	64.3	69.4
	Ora.	71.8	71.0	72.8	76.6	63.8	62.9	65.3	70.8
Caltech101	Seq.	85.1	96.1	97.2	96.3	81.9	91.9	94.6	94.5
	Cum.	86.5	96.0	97.2	97.0	83.9	93.0	95.5	95.0
	Ora.	86.5	96.3	97.2	97.0	84.1	93.3	95.7	95.4
OxfordPets	Seq.	83.6	84.3	91.2	90.7	83.6	71.8	87.8	87.5
	Cum.	90.0	86.4	91.7	92.1	85.4	69.0	87.2	87.6
	Ora.	92.8	87.6	92.8	92.8	85.0	70.3	88.7	89.3
StanfordCars	Seq.	80.8	87.3	91.1	86.4	79.1	83.0	89.9	86.0
	Cum.	87.7	89.5	92.3	88.9	85.2	85.0	91.0	88.9
	Ora.	88.1	89.8	92.8	91.3	83.8	85.1	91.0	89.0
Flowers102	Seq.	69.6	97.1	99.0	96.0	57.5	95.0	98.4	94.2
	Cum.	72.1	97.7	99.4	97.0	64.8	95.3	98.3	94.4
	Ora.	77.4	97.8	99.4	97.3	66.6	95.6	98.4	96.8
Food101	Seq.	84.4	85.2	90.1	88.0	80.6	81.2	88.8	86.8
	Cum.	88.7	87.8	92.4	91.4	84.1	83.5	89.9	88.2
	Ora.	89.9	89.0	93.0	91.7	85.2	83.6	90.8	89.5
FGVCAircraft	Seq.	19.1	50.5	61.7	43.9	8.3	38.3	50.5	30.1
	Cum.	26.8	54.2	65.1	49.2	11.8	38.8	52.9	33.2
	Ora.	25.4	57.3	66.5	48.6	13.7	40.4	55.6	30.8
SUN397	Seq.	64.1	71.5	75.9	73.0	62.7	69.6	75.0	71.9
	Cum.	67.9	73.4	77.6	75.7	65.7	70.5	75.6	72.3
	Ora.	67.0	74.3	78.6	76.5	54.1	71.2	76.5	74.5
DTD	Seq.	52.0	71.6	79.2	72.6	48.0	68.6	78.7	74.6
	Cum.	58.2	73.4	81.0	74.7	51.1	69.3	79.0	76.1
	Ora.	58.8	73.4	81.9	77.8	53.5	68.0	79.3	76.0
EuroSAT	Seq.	55.4	84.0	96.7	95.2	46.9	78.7	96.2	93.7
	Cum.	53.6	85.1	96.8	95.4	46.6	74.4	96.2	93.0
	Ora.	52.9	86.0	96.8	95.3	31.5	71.8	95.7	94.6
UCF101	Seq.	59.0	76.7	83.1	79.7	53.7	72.7	80.8	76.8
	Cum.	64.1	79.4	85.2	80.7	55.8	74.2	81.1	77.3
	Ora.	67.6	80.6	85.6	83.9	59.8	73.3	81.7	77.9
Mean	Seq.	65.3	79.3	85.0	81.3	60.1	73.9	82.2	78.5
	Cum.	69.7	81.1	86.4	83.4	63.3	74.1	82.8	79.6
	Ora.	70.7	82.1	87.0	84.4	61.9	74.1	83.5	80.4

D ONE-YEAR DECLINES ALONG THE TRAJECTORY

This section supports Section 4. Along the Cumulative trajectory each model is trained on a superset of the previous one's data, so none of the declines recorded here can come from data the model has stopped seeing. Table 12 identifies the ten steps of Figure 1 (b). Table 13 shows the count is not an artifact of the one-point threshold, Table 14 rules out sampling noise of the test sets, and Tables 15 and 16 rule out fitting noise of the image-side protocols.

E GAPS BETWEEN TRAINING STRATEGIES

This section supports Section 5. Table 17 gives the per-dataset gaps behind every average in Table 3 of the main paper, together with the significance tests, and Figure 4 plots them against the zero-shot gaps on the same checkpoints. Table 18 tests whether any single dataset drives those averages. Table 19 follows the replay gap over time and supports Figure 3. Table 20 repeats the replay comparison with the image-side protocols interpolated to the accuracy level of the zero-shot head. Table 21 reads the same gap as a delay, a unit that saturates over a six-year run.

Table 10: **Zero-shot accuracy of one frozen checkpoint under five prompt configurations.** The Cumulative 2022 checkpoint under bestpool; the weights are identical in every column. Local names are the class names shipped with each dataset, CLIP names those of Radford et al. (2021), which rewrite ambiguous labels and add a category word. Generic is the single template `a photo of a {}`, CLIP template the single template of Radford et al. (2021), and CLIP ensemble the mean over their full per-dataset template set. The last column is the configuration used everywhere else in this paper. Range is the largest minus the smallest value in the row. Ranges are computed from unrounded accuracies.

	Generic template		CLIP template		CLIP ensemble	
Dataset	local names	CLIP names	local names	CLIP names	CLIP names	Range
ImageNet	69.5	69.5	70.0	70.0	70.8	1.4
Caltech101	83.7	86.7	84.5	86.4	86.5	3.1
OxfordPets	88.6	88.6	90.0	90.0	90.0	1.4
StanfordCars	87.0	87.0	87.2	87.2	87.7	0.7
Flowers102	71.9	72.1	72.0	72.1	72.1	0.2
Food101	87.4	88.9	87.5	88.6	88.7	1.5
FGVCAircraft	25.2	25.2	25.7	25.7	26.8	1.6
SUN397	65.6	67.5	67.2	68.4	67.9	2.7
DTD	53.1	53.1	56.9	56.9	58.2	5.1
EuroSAT	24.3	38.8	52.4	55.6	53.6	31.3
UCF101	59.3	59.2	62.6	62.3	64.1	4.9
Mean	65.1	67.0	68.7	69.4	69.7	4.6
Median range						1.6

Table 11: **The Oracle minus Cumulative gap under five prompt configurations.** Difference in zero-shot accuracy (points) between the two final checkpoints, bestpool / basic, with the columns of Table 10. Sign changes counts datasets whose difference does not keep the same sign across all five configurations; range is the largest minus the smallest of the five differences. Every column evaluates the same two checkpoints. Ranges are computed from unrounded accuracies.

	Generic template		CLIP template		CLIP ensemble	
Dataset	local names	CLIP names	local names	CLIP names	CLIP names	Range
ImageNet	+1.2 / −2.4	+1.2 / −2.4	+0.9 / −1.0	+0.9 / −1.0	+1.0 / +1.5	0.3 / 3.9
Caltech101	−0.6 / −4.5	−0.4 / −1.4	−1.9 / +0.5	−0.2 / −3.2	−0.1 / +0.2	1.8 / 5.0
OxfordPets	+3.5 / −2.3	+3.5 / −2.3	+2.8 / −0.4	+2.8 / −0.4	+2.8 / −0.4	0.6 / 1.9
StanfordCars	+0.4 / −4.0	+0.4 / −4.0	+0.5 / −3.3	+0.5 / −3.3	+0.3 / −1.4	0.2 / 2.6
Flowers102	+2.2 / −3.5	+2.4 / −2.9	+5.2 / +0.9	+5.3 / +1.8	+5.3 / +1.8	3.1 / 5.2
Food101	+1.4 / −1.6	+1.0 / −0.5	+1.3 / +1.4	+1.3 / +1.1	+1.3 / +1.1	0.4 / 3.0
FGVCAircraft	−1.8 / −2.6	−1.8 / −2.6	−0.9 / +1.2	−0.9 / +1.2	−1.4 / +1.9	0.9 / 4.5
SUN397	−0.4 / −8.1	−1.1 / −7.0	−1.1 / −12.4	−0.8 / −10.7	−0.9 / −11.6	0.7 / 5.4
DTD	+1.0 / +1.4	+1.0 / +1.4	+1.1 / +2.2	+1.1 / +2.2	+0.6 / +2.5	0.5 / 1.1
EuroSAT	+23.5 / −2.0	+9.7 / +2.0	−1.3 / +10.1	−3.6 / −16.6	−0.8 / −15.1	27.1 / 26.6
UCF101	+5.4 / +0.8	+5.4 / +1.0	+3.0 / +4.5	+3.2 / +4.1	+3.5 / +4.0	2.5 / 3.7
Mean over datasets	+3.2 / −2.6	+1.9 / −1.7	+0.9 / +0.3	+0.9 / −2.2	+1.1 / −1.4	3.5 / 5.7
Oracle leads (of 11)	8 / 2	8 / 3	7 / 7	7 / 5	7 / 7	
Sign changes across configurations		1 / 6 (of 11)				

F Robustness under distribution shift

The benchmark reports accuracy on six natural ImageNet shifts but does not ask whether replay improves robustness beyond accuracy, and ImageNet-C is not in its suite. This section takes up the question raised in Section 6.3. Two things stand in the way of reading an excess gain on a shifted set as added robustness: the zero-shot head depends on class-name embeddings, and shifted accuracy rises with ImageNet accuracy, not necessarily one for one. We therefore repeat the comparison with a linear probe trained on each checkpoint's clean ImageNet features (Table 22) and summarize

Table 12: **Every one-year zero-shot decline of at least one point along the Cumulative trajectory.** Ten of the 132 steps (11 datasets $\times$ 6 steps $\times$ 2 filtering regimes). Δ is the change in accuracy (points) from one time step to the next, measured on the same pair of frozen checkpoints by all three protocols. The training pool grows at every one of these steps, so none of the declines can come from data the model has stopped seeing. ZS is zero-shot classification with the CLIP's class names and template, NCM is nearest class mean on the image features, and LP is a linear probe on the full training split.

Filter	Dataset	Step	ΔZS	ΔNCM	ΔLP
Bestpool	EuroSAT	2016 $\rightarrow$ 2017	-1.7	-0.8	-0.1
	Flowers102	2021 $\rightarrow$ 2022	-1.4	$+0.0$	$+0.1$
	EuroSAT	2021 $\rightarrow$ 2022	-3.8	-1.1	$+0.1$
Basic	Flowers102	2017 $\rightarrow$ 2018	-2.0	$+0.1$	-0.1
	SUN397	2017 $\rightarrow$ 2018	-5.1	$+1.5$	$+1.1$
	DTD	2017 $\rightarrow$ 2018	-1.0	$+0.3$	$+0.1$
	EuroSAT	2017 $\rightarrow$ 2018	-1.4	$+0.6$	$+0.3$
	UCF101	2017 $\rightarrow$ 2018	-1.5	$+0.8$	-0.6
	Flowers102	2020 $\rightarrow$ 2021	-1.8	-0.3	$+0.2$
	EuroSAT	2021 $\rightarrow$ 2022	-2.6	$+0.4$	$+0.2$
Steps where the protocol falls by at least one point			10	1	0
Steps where the protocol rises			0	6	7

Table 13: **One-year steps of the Cumulative trajectory, pooled and by filtering regime.** A step counts as a zero-shot decline if $\Delta\text{ZS} \leq -\delta$. The two image-side columns count, *among those same steps*, how many also decline by at least δ; the last column how many of them show nearest class mean moving up while zero-shot moves down. The pool grows at every step, so none of these declines can come from data the model has stopped seeing. Pooled over all 132 steps, the mean absolute one-year change is 1.95, 0.90 and 0.70 points under the three protocols, and the range is -5.1 to $+19.4$, -3.3 to $+5.8$, and -1.3 to $+4.3$.

		Steps			
Group	δ	$\Delta\text{ZS} \leq -\delta$	also $\Delta\text{NCM} \leq -\delta$	also $\Delta\text{LP} \leq -\delta$	also $\Delta\text{NCM} > 0$
Pooled (132)	0.5	21	3	2	13
	1.0	10	1	0	6
	2.0	4	0	0	3
Bestpool (66)	0.5	9	2	1	6
	1.0	3	1	0	0
	2.0	1	0	0	0
Basic (66)	0.5	12	1	1	7
	1.0	7	0	0	6
	2.0	3	0	0	3

SD of one-year change	ZS	NCM	LP
pooled	3.10	1.23	1.01
bestpool	3.11	1.05	0.90
basic	3.11	1.39	1.12

each step by its break-even slope (Table 23), the trend steep enough to explain the step's shifted gain from its ImageNet gain alone; under any baseline flatter than this, the step raises effective robustness (Taori et al., 2020).

The probe supports a robustness benefit only under basic. There, replay has the smaller ImageNet gain (0.4 against 1.0 points) but the larger ImageNet-C gain (1.7 against 0.7), giving break-even slopes of 5.3 and 0.8, and the zero-shot head orders the two steps the same way (1.3 against 0.2). That ratio is unstable, since its denominator is a gain of 0.4 points: recomputing it from the three-

Table 14: Do the zero-shot drops of Table 13 exceed the sampling noise of the test sets? For each of the 66 adjacent Cumulative checkpoint pairs in each regime and each protocol, McNemar's exact test on the same test set and a paired bootstrap 95% CI for the accuracy difference. The test rules out sampling noise; the cumulative design, where the pool only grows, is what rules out a loss of training data.

		Significant decreases (of 66)			Median CI width (pp)		
Filter	Test	ZS	NCM	LP	ZS	NCM	LP
Bestpool	McNemar $p < 0.05$	6	2	0	1.42	1.11	0.87
	bootstrap CI excl. 0	6	2	0	1.42	1.11	0.87
Basic	McNemar $p < 0.05$	6	5	1	1.58	1.50	1.04
	bootstrap CI excl. 0	6	5	2	1.58	1.50	1.04

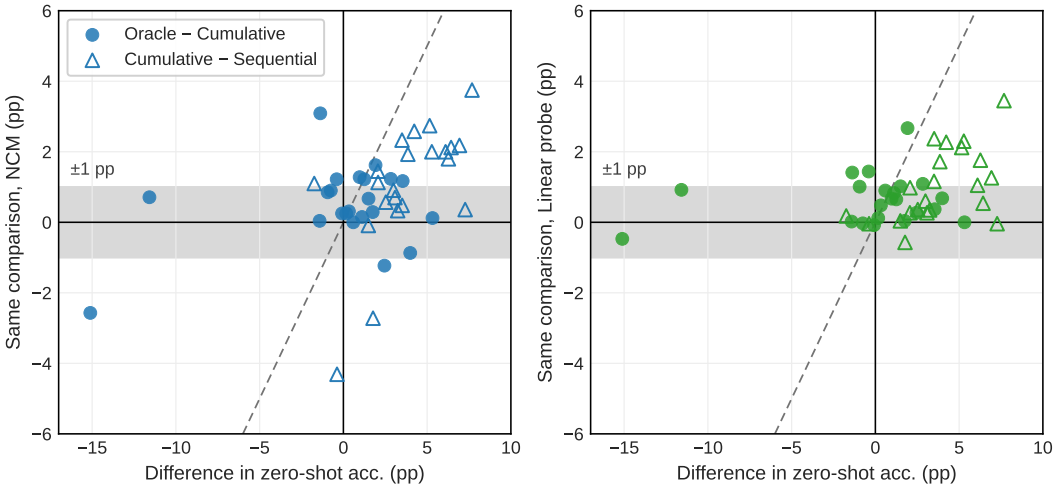


Figure 4: **The gap between training strategies, measured in two ways.** Each point is one dataset under one filter. Horizontal: the difference between two final checkpoints in zero-shot accuracy. Vertical: the same difference measured by nearest class mean (left) or a linear probe (right). Filled: O−C; open: C−S. If the protocols measured the same quantity, the points would lie on $y = x$. Shaded band: ± 1 point.

decimal accuracies of Table 22 gives 4.4 rather than 5.3. Under bestpool, replay's larger excess reflects its larger ImageNet gain (2.3 against 0.7 points), and the two break-even slopes, 1.7 and 2.0, sit at the reference slope of 1.7, so neither step clearly changes effective robustness. On natural shifts neither step has a consistent advantage, with break-even slopes of 1.2 and 2.0 under bestpool and 2.4 and 1.0 under basic against a reference slope of 1.1. Replay reliably improves accuracy; on corruptions it beats the trend only under basic filtering, and on natural shifts under neither.

G Repair on the text side

Sections 4 and 5 separate the image features from the text side by removing the text encoder. This section asks the complementary question: how much of what the zero-shot head misses can be recovered by moving the text side instead, and at what cost in labels. CoOp (Zhou et al., 2022b) freezes both towers and re-learns only the context vectors of the prompt, so the image features, and with them the NCM and LP columns, are unchanged by construction.

Table 9 gives the full-data result. The comparison of interest is not CoOp's accuracy, which is expected to exceed the zero-shot head and in fact stays below the linear probe on every checkpoint, but what optimizing the prompt does to the gaps between checkpoints. Under basic filtering the Oracle versus Cumulative difference is -1.4 points under the zero-shot evaluation, $+0.8$ under a learned one and $+0.7$ under the linear probe; the replay gap is 3.2, 1.1 and 0.6. Under bestpool the

Table 15: k-shot nearest-class-mean accuracy (%), mean over three seeds. Within a seed, all six checkpoints use the same k-shot subset (paired design). *full*: all training images. $\bar{s}$: across-seed standard deviation averaged over the eleven datasets.

k	Model	IN	Cal	Pet	Car	Flo	Food	Air	SUN	DTD	ESAT	UCF	Mean	Med.	$\bar{s}$
Bestpool															
1	Seq.	28.6	79.4	45.7	45.5	78.8	44.6	19.1	37.1	42.4	52.9	47.5	47.4	45.5	1.44
	Cum.	31.3	82.0	47.9	50.1	80.2	48.3	20.3	40.0	42.8	58.3	52.3	50.3	48.3	1.56
	Ora.	33.2	83.6	50.0	51.8	83.4	51.7	21.4	41.2	42.1	58.4	53.3	51.8	51.7	1.68
4	Seq.	51.3	93.1	70.9	72.2	93.9	71.6	34.7	59.6	62.3	74.9	67.3	68.3	70.9	1.37
	Cum.	54.2	93.7	73.9	76.3	94.8	75.7	36.8	62.1	63.3	79.4	71.2	71.0	73.9	0.88
	Ora.	55.9	93.7	75.4	77.5	95.9	78.4	38.4	63.4	63.5	78.4	72.9	72.1	75.4	1.14
8	Seq.	58.4	94.5	78.4	79.8	95.7	78.6	40.2	65.9	68.0	80.1	71.6	73.8	78.4	0.60
	Cum.	61.1	94.6	81.5	82.8	96.4	82.0	42.9	68.0	69.5	82.0	75.1	76.0	81.5	0.58
	Ora.	62.7	95.1	82.4	84.1	97.0	84.2	44.6	69.3	68.9	82.3	77.3	77.1	82.3	0.78
16	Seq.	62.8	95.2	81.9	84.7	96.5	81.7	44.6	69.3	69.3	82.0	73.7	76.5	81.7	0.41
	Cum.	65.2	95.2	84.1	87.1	97.5	84.7	47.5	71.3	71.0	83.6	76.9	78.6	83.6	0.33
	Ora.	66.6	95.5	85.9	87.9	97.8	86.5	49.5	72.3	71.5	85.4	78.6	79.8	85.4	0.36
full	Seq.	67.7	96.1	84.3	87.3	97.1	85.2	50.5	71.5	71.6	84.0	76.7	79.3	84.0	0.00
	Cum.	69.7	96.0	86.4	89.5	97.7	87.8	54.2	73.4	73.4	85.1	79.4	81.1	85.1	0.00
	Ora.	71.0	96.3	87.6	89.8	97.8	89.0	57.3	74.3	73.4	86.0	80.6	82.1	86.0	0.00
Basic															
1	Seq.	22.7	68.6	32.7	36.1	65.0	30.2	12.4	31.9	34.1	48.6	41.5	38.5	34.1	1.69
	Cum.	23.1	69.5	32.0	38.9	66.4	33.0	12.4	32.9	34.3	44.9	44.3	39.2	34.3	2.16
	Ora.	23.9	71.4	31.8	40.5	65.0	28.7	14.1	34.6	34.5	43.3	42.9	39.2	34.6	2.12
4	Seq.	43.9	86.5	56.2	64.6	87.9	58.7	23.9	56.0	57.0	68.1	63.0	60.5	58.7	0.98
	Cum.	44.3	87.5	53.8	69.0	88.7	61.8	23.5	57.0	57.9	61.8	64.5	60.9	61.8	1.18
	Ora.	45.3	88.1	54.5	69.9	87.9	57.9	26.1	58.6	56.8	62.4	63.0	61.0	58.6	1.11
8	Seq.	51.3	89.5	64.1	73.8	92.0	68.2	28.5	63.3	63.6	74.7	67.6	67.0	67.6	0.79
	Cum.	51.7	90.6	61.4	77.2	92.7	71.6	28.4	64.1	64.7	69.9	68.7	67.4	68.7	0.69
	Ora.	52.6	90.9	62.8	77.5	92.6	69.0	30.7	65.4	62.7	68.0	67.5	67.3	67.5	0.68
16	Seq.	56.1	90.7	69.3	79.7	93.9	73.5	32.3	67.3	66.4	76.6	70.0	70.5	70.0	0.59
	Cum.	56.4	91.5	66.6	82.3	94.8	76.8	32.3	68.0	66.5	71.6	70.6	70.7	70.6	0.70
	Ora.	57.3	91.9	67.4	82.5	94.6	74.7	34.5	69.1	65.1	69.4	70.5	70.6	69.4	0.59
full	Seq.	61.9	91.9	71.8	83.0	95.0	81.2	38.3	69.6	68.6	78.7	72.7	73.9	72.7	0.00
	Cum.	62.2	93.0	69.0	85.0	95.3	83.5	38.8	70.5	69.3	74.4	74.2	74.1	74.2	0.00
	Ora.	62.9	93.3	70.3	85.1	95.6	83.6	40.4	71.2	68.0	71.8	73.3	74.1	71.8	0.00

Table 16: k-shot linear-probe accuracy (%), mean over three seeds, with the same paired k-shot subsets as Table 15. *full*: probe fit on the entire training split.

k	Model	IN	Cal	Pet	Car	Flo	Food	Air	SUN	DTD	ESAT	UCF	Mean	Med.	$\bar{s}$
Bestpool															
1	Seq.	30.7	82.7	46.8	47.0	79.7	47.3	19.4	38.8	43.5	56.5	49.7	49.3	47.0	1.96
	Cum.	33.0	84.4	49.8	52.4	81.2	50.9	21.3	41.5	43.8	60.1	54.1	52.0	50.9	2.04
	Ora.	34.8	86.0	52.1	53.9	83.5	53.9	21.8	42.7	44.0	61.2	55.7	53.6	53.9	1.73
4	Seq.	53.8	93.8	72.9	74.1	94.8	73.1	35.9	61.1	63.1	82.9	69.3	70.4	72.9	1.20
	Cum.	56.8	94.3	76.5	78.6	95.9	77.7	37.8	64.0	65.5	84.6	73.4	73.2	76.5	0.97
	Ora.	58.7	94.5	74.7	80.0	97.0	80.1	39.2	65.1	63.8	82.7	74.6	73.7	74.7	1.07
8	Seq.	61.2	95.6	81.0	83.0	97.3	80.0	43.1	68.1	71.0	87.4	74.1	76.5	80.0	0.52
	Cum.	64.3	96.2	83.4	86.1	97.7	83.7	45.0	70.2	72.5	87.8	78.2	78.7	83.4	0.55
	Ora.	65.8	96.1	83.7	87.0	98.2	85.4	46.8	71.5	72.6	87.7	79.5	79.5	83.7	0.52
16	Seq.	66.6	96.6	84.0	88.0	98.2	83.5	49.8	72.2	74.7	89.6	78.3	80.1	83.5	0.41
	Cum.	68.8	96.6	86.8	90.3	98.8	86.8	52.7	74.3	76.1	90.7	81.4	82.1	86.8	0.25
	Ora.	70.5	96.5	88.0	90.9	99.0	88.0	54.1	75.3	76.9	90.5	82.3	82.9	88.0	0.34
full	Seq.	69.8	97.2	91.2	91.1	99.0	90.1	61.7	75.9	79.2	96.7	83.1	85.0	90.1	0.00
	Cum.	72.1	97.2	91.7	92.3	99.4	92.4	65.1	77.6	81.0	96.8	85.2	86.4	91.7	0.00
	Ora.	72.8	97.2	92.8	92.8	99.4	93.0	66.5	78.6	81.9	96.8	85.6	87.0	92.8	0.00
Basic															
1	Seq.	25.6	72.8	34.7	39.0	66.8	34.9	13.3	34.3	35.5	52.9	44.4	41.3	35.5	1.95
	Cum.	26.2	74.5	35.5	43.3	67.5	36.3	12.7	35.5	36.2	46.9	45.8	41.9	36.3	1.99
	Ora.	26.7	76.3	35.3	44.8	66.8	33.1	14.4	36.3	36.6	41.4	44.8	41.5	36.6	1.90
4	Seq.	47.4	89.1	60.9	69.8	90.2	65.9	25.1	58.0	59.5	79.0	65.1	64.6	65.1	0.95
	Cum.	48.4	89.8	59.8	74.1	90.9	68.4	25.5	59.3	61.4	72.3	67.0	65.2	67.0	1.08
	Ora.	49.6	90.2	61.6	75.0	91.4	68.2	27.8	60.8	61.1	70.6	66.4	65.7	66.4	1.61
8	Seq.	55.7	91.9	72.1	79.4	94.7	74.3	31.2	65.6	67.9	83.6	71.3	71.6	72.1	0.80
	Cum.	56.5	92.6	71.0	82.3	94.8	77.2	32.6	67.0	68.6	79.5	72.4	72.2	72.4	0.73
	Ora.	57.7	93.1	73.0	83.1	95.1	77.9	33.5	68.2	68.7	80.5	72.3	73.0	73.0	0.82
16	Seq.	61.5	93.7	77.6	85.6	96.8	79.4	38.3	70.7	72.8	85.5	75.8	76.1	77.6	0.83
	Cum.	62.1	94.0	76.2	87.8	97.1	81.8	39.6	71.9	73.7	83.6	76.0	76.7	76.2	0.49
	Ora.	63.4	94.2	79.0	88.0	97.4	82.3	41.1	72.9	73.0	85.1	76.5	77.5	79.0	0.61
full	Seq.	63.9	94.6	87.8	89.9	98.4	88.8	50.5	75.0	78.7	96.2	80.8	82.2	87.8	0.00
	Cum.	64.3	95.5	87.2	91.0	98.3	89.9	52.9	75.6	79.0	96.2	81.1	82.8	87.2	0.00
	Ora.	65.3	95.7	88.7	91.0	98.4	90.8	55.6	76.5	79.3	95.7	81.7	83.5	88.7	0.00

corresponding values are $+1.1$, $+1.0$ and $+0.6$, and 4.4, 2.2 and 1.5. A prompt fitted to the frozen features recovers most of the image-side answer, which is what a gap that lives in the alignment rather than in the features would predict.

Table 24 asks how many labels this takes, using the same k-shot subsets as Tables 15 and 16. One image per class already beats the zero-shot head on all six checkpoints, by 1.9 to 4.7 points. Overtaking nearest class mean on the full training split takes more: eight shots per class for Cumulative and Oracle under basic, sixteen for Cumulative and Oracle under bestpool and for Sequential under basic, and the full split for Sequential under bestpool, where sixteen shots fall 0.2 points short. The linear probe on the full split stays ahead of full-data CoOp on every checkpoint.

Table 17: **Per-dataset gaps at 2022 under three protocols of the same frozen features** (pp). Removing the text head shrinks the gap, collapses its spread across datasets, and makes its sign consistent. p is the exact two-sided Wilcoxon signed-rank test over the eleven datasets. Differences and the means in Table 3 are computed from unrounded accuracies and rounded once, so they may differ by 0.1 from differences of the rounded accuracies in Tables 6–9.

	Bestpool			Basic		
Dataset	ZS	NCM	LP	ZS	NCM	LP
(a) $\Delta_{\mathrm{C-S}}$: Cumulative − Sequential						
ImageNet	+5.3	+2.0	+2.3	+3.2	+0.3	+0.4
Caltech101	+1.5	−0.1	+0.0	+2.1	+1.1	+1.0
OxfordPets	+6.4	+2.1	+0.5	+1.8	−2.7	−0.6
StanfordCars	+6.9	+2.2	+1.3	+6.1	+2.0	+1.1
Flowers102	+2.5	+0.6	+0.4	+7.3	+0.4	+0.0
Food101	+4.2	+2.6	+2.3	+3.5	+2.3	+1.2
FGVCAircraft	+7.7	+3.8	+3.5	+3.5	+0.5	+2.4
SUN397	+3.9	+1.9	+1.7	+3.0	+0.9	+0.6
DTD	+6.3	+1.8	+1.8	+3.1	+0.7	+0.3
EuroSAT	−1.7	+1.1	+0.2	−0.4	−4.3	+0.0
UCF101	+5.2	+2.7	+2.1	+2.1	+1.5	+0.3
Median	+5.2	+2.0	+1.7	+3.1	+0.7	+0.4
SD	2.76	1.05	1.08	2.07	2.00	0.79
#> 0 (of 11)	10	10	11	10	9	8
p	0.003	0.002	0.001	0.002	0.320	0.029
(b) $\Delta_{\mathrm{O-C}}$: Oracle − Cumulative						
ImageNet	+1.0	+1.3	+0.7	+1.5	+0.7	+1.0
Caltech101	−0.1	+0.3	−0.1	+0.2	+0.3	+0.1
OxfordPets	+2.8	+1.2	+1.1	−0.4	+1.2	+1.4
StanfordCars	+0.3	+0.3	+0.5	−1.4	+0.0	+0.0
Flowers102	+5.3	+0.1	+0.0	+1.8	+0.3	+0.0
Food101	+1.3	+1.2	+0.7	+1.1	+0.2	+0.8
FGVCAircraft	−1.4	+3.1	+1.4	+1.9	+1.6	+2.7
SUN397	−0.9	+0.9	+1.0	−11.6	+0.7	+0.9
DTD	+0.6	+0.0	+0.9	+2.5	−1.2	+0.3
EuroSAT	−0.8	+0.9	+0.0	−15.1	−2.6	−0.5
UCF101	+3.5	+1.2	+0.4	+4.0	−0.9	+0.7
Median	+0.6	+0.9	+0.7	+1.1	+0.3	+0.7
SD	2.07	0.86	0.49	6.12	1.19	0.86
#> 0 (of 11)	7	10	8	7	8	10
p	0.206	0.002	0.010	0.638	0.638	0.010
(c) $\Delta_{\mathrm{O-S}}$: Oracle − Sequential						
ImageNet	+6.3	+3.3	+3.0	+4.7	+1.0	+1.4
Caltech101	+1.4	+0.2	+0.0	+2.2	+1.4	+1.1
OxfordPets	+9.3	+3.4	+1.6	+1.4	−1.5	+0.9
StanfordCars	+7.3	+2.5	+1.7	+4.7	+2.0	+1.1
Flowers102	+7.9	+0.7	+0.4	+9.0	+0.7	+0.0
Food101	+5.5	+3.8	+2.9	+4.6	+2.5	+2.0
FGVCAircraft	+6.3	+6.8	+4.9	+5.4	+2.1	+5.0
SUN397	+2.9	+2.8	+2.7	−8.6	+1.6	+1.5
DTD	+6.9	+1.8	+2.7	+5.5	−0.5	+0.5
EuroSAT	−2.5	+2.0	+0.2	−15.5	−6.9	−0.5
UCF101	+8.7	+3.9	+2.5	+6.1	+0.6	+1.0
Median	+6.3	+2.8	+2.5	+4.7	+1.0	+1.1
SD	3.51	1.80	1.47	7.28	2.65	1.43
#> 0 (of 11)	10	11	10	9	8	9
p	0.003	0.001	0.002	0.278	0.206	0.004

Exact HL intervals for (a), bestpool: ZS $[+2.53, +6.27]$, NCM $[+1.19, +2.46]$, LP $[+0.72, +2.20]$; basic: ZS $[+1.92, +4.67]$, NCM $[-1.18, +1.41]$, LP $[+0.12, +1.11]$. For (b), basic: ZS $[-6.33, +1.92]$, LP $[+0.12, +1.23]$.

Table 18: **Leave-one-dataset-out sensitivity of the average gap** (pp). Range is the interval spanned by the eleven averages obtained by dropping one dataset; Width is its length. Leverage names the dataset that pulls the average furthest from the value it would take without it, and by how much, $(x_i - \text{mean})/10$. The last two columns repeat the exercise for the median.

Filter	Protocol	Mean	Range	Width	Leverage	Median	Median range
(a) $\Delta_{\mathrm{C-S}}$: Cumulative − Sequential							
	ZS	$+4.4$	$[+4.04, +4.98]$	0.94	EuroSAT -0.61	$+5.2$	$[+4.69, +5.22]$
Bestpool	NCM	$+1.9$	$[+1.69, +2.08]$	0.38	Caltech -0.20	$+2.0$	$[+1.97, +2.06]$
	LP	$+1.5$	$[+1.26, +1.60]$	0.34	Aircraft $+0.20$	$+1.7$	$[+1.49, +1.74]$
	ZS	$+3.2$	$[+2.79, +3.56]$	0.77	Flowers $+0.41$	$+3.1$	$[+3.04, +3.16]$
Basic	NCM	$+0.2$	$[+0.03, +0.70]$	0.66	EuroSAT -0.46	$+0.7$	$[+0.59, +0.81]$
	LP	$+0.6$	$[+0.40, +0.70]$	0.29	Aircraft $+0.18$	$+0.4$	$[+0.31, +0.48]$
(b) $\Delta_{\mathrm{O-C}}$: Oracle − Cumulative							
	ZS	$+1.1$	$[+0.64, +1.31]$	0.67	Flowers $+0.43$	$+0.6$	$[+0.46, +0.78]$
Bestpool	NCM	$+0.9$	$[+0.73, +1.04]$	0.31	Aircraft $+0.21$	$+0.9$	$[+0.88, +1.04]$
	LP	$+0.6$	$[+0.51, +0.66]$	0.15	Aircraft $+0.08$	$+0.7$	$[+0.57, +0.66]$
	ZS	-1.4	$[-1.96, -0.05]$	1.91	EuroSAT -1.37	$+1.1$	$[+0.64, +1.31]$
Basic	NCM	$+0.0$	$[-0.13, +0.29]$	0.42	EuroSAT -0.26	$+0.3$	$[+0.20, +0.27]$
	LP	$+0.7$	$[+0.49, +0.80]$	0.31	Aircraft $+0.20$	$+0.7$	$[+0.48, +0.76]$

Five of the eleven leave-one-out averages change sign for $\Delta_{\mathrm{O-C}}$ under NCM / basic, where the average is itself $+0.03$; none do in any other row. Under ZS / basic the average stays negative in all eleven, but dropping EuroSAT alone brings it to -0.05, and dropping EuroSAT and SUN397 together makes it $+1.23$.

Table 19: **Median replay gap $\Delta_{\mathrm{C-S}}$ over time, per protocol.** This table holds the values plotted in Figure 3. $\delta = \Delta\text{ZS} - \Delta\text{LP}$ is computed per dataset before taking the median. Slope is the median over datasets of the per-dataset least-squares slope of the gap on time, in points per year. A cost of dropping replay that lives at the protocol rather than in the features predicts a positive zero-shot slope together with near-zero NCM and LP slopes, which is what the last row shows.

	Bestpool				Basic			
Year	ZS	NCM	LP	δ	ZS	NCM	LP	δ
2017	$+1.0$	$+0.2$	$+0.3$	$+0.8$	$+0.6$	$+0.1$	$+0.1$	$+0.6$
2018	$+2.9$	$+0.6$	$+0.6$	$+1.6$	$+1.4$	$+0.1$	$+0.5$	$+0.1$
2019	$+2.0$	$+0.6$	$+0.4$	$+1.6$	$+2.0$	$+0.5$	$+0.6$	$+1.3$
2020	$+3.3$	$+0.7$	$+0.8$	$+2.5$	$+2.2$	$+0.6$	$+0.7$	$+1.7$
2021	$+2.4$	$+0.7$	$+0.9$	$+1.7$	$+3.2$	$+0.7$	$+0.4$	$+2.8$
2022	$+5.2$	$+2.0$	$+1.7$	$+3.0$	$+3.1$	$+0.7$	$+0.4$	$+2.3$
Slope (pp/yr)	$+0.6$	$+0.2$	$+0.2$	$+0.4$	$+0.5$	$+0.1$	$+0.0$	$+0.4$

Table 20: Accuracy-matched comparison of the replay gap (Cumulative − Sequential, pp). For each dataset and protocol, $k^\star$ is the shot count at which the mean k-shot accuracy of the two checkpoints equals their mean zero-shot accuracy; it is obtained by linear interpolation in $\log_2 k$, through the sample budgets $k \in \{1, 4, 8, 16, full\}$, with $k_{\text{full}} = n_{\text{train}}/n_{\text{classes}}$, and $\Delta@k^\star$ is the gap interpolated at $k^\star$. Entries < 1 or >full mean that the image protocol already exceeds the zero-shot level with one shot, or never reaches it; those 14 of the 44 dataset-protocol pairs cannot be matched and are excluded from the summary rows, which is the one place in the paper where datasets are dropped from an average. The summary rows compare Δ_{ZS} with $\Delta@k^\star$ on the matched pairs of each column (exact two-sided Wilcoxon signed-rank test; at $n = 6$, $p = 0.031$ is the smallest attainable value and is reached only when every dataset agrees). If zero-shot inflates the gap, $\Delta@k^\star < \Delta_{\text{ZS}}$ at matched accuracy.

	Bestpool					Basic				
		NCM		LP			NCM		LP	
Dataset	Δ_{ZS}	$k^\star$	$\Delta@k^\star$	$k^\star$	$\Delta@k^\star$	Δ_{ZS}	$k^\star$	$\Delta@k^\star$	$k^\star$	$\Delta@k^\star$
ImageNet	+5.3	805.2	+2.0	31.6	+2.2	+3.2	459.9	+0.3	14.0	+0.7
Caltech101	+1.5	1.7	+1.8	1.3	+1.5	+2.1	2.9	+1.0	2.2	+1.1
OxfordPets	+6.4	>full	–	24.3	+2.3	+1.8	>full	–	59.0	−0.8
StanfordCars	+6.9	12.5	+2.6	7.8	+3.1	+6.1	21.3	+2.4	9.4	+2.8
Flowers102	+2.5	<1	–	<1	–	+7.3	<1	–	<1	–
Food101	+4.2	>full	–	38.1	+3.1	+3.5	>full	–	35.1	+2.2
FGVCAircraft	+7.7	1.3	+1.5	1.2	+1.9	+3.5	<1	–	<1	–
SUN397	+3.9	7.2	+2.2	5.8	+2.5	+3.0	8.8	+0.8	6.6	+1.4
DTD	+6.3	2.4	+0.8	2.2	+1.5	+3.1	2.5	+0.7	2.2	+1.3
EuroSAT	−1.7	<1	–	<1	–	−0.4	<1	–	<1	–
UCF101	+5.2	2.3	+4.3	2.0	+4.2	+2.1	2.2	+2.1	1.9	+1.7
Median (all datasets)	+5.2					+3.1				
# matched		7		9			6		8	
# $\Delta@k^\star < \Delta_{\text{ZS}}$		6		9			6		8	
Median $\Delta_{\text{ZS}} - \Delta@k^\star$		+3.24		+3.06			+2.30		+1.71	
Wilcoxon p		0.031		0.004			0.031		0.008	

Table 21: **The replay gap read as a delay.** For each dataset, $2022 - t^\star$, where $t^\star$ is the year at which a monotone fit of the Cumulative trajectory, linearly interpolated between years, first reaches the accuracy Sequential ends with in 2022; censored to $[0, 6]$ years and reported as the median over datasets. # is the number of datasets with a positive lag. The unit saturates: the trajectory runs for only six years and flattens near the end, so under bestpool all three protocols fall within half a year of each other and none is far from the ceiling. We therefore read the gap as a fraction of Cumulative's own progress (Table 4) rather than as a delay.

	Bestpool		Basic	
Protocol	Lag (yr)	#	Lag (yr)	#
ZS	5.5	11	4.1	11
NCM-full	5.0	11	3.7	9
LP-full	5.1	11	3.3	8

Table 22: **Accuracy of the six final checkpoints under two heads.** ZS is the zero-shot head. LP is a linear probe trained separately for each checkpoint on its clean ImageNet features with multinomial logistic regression, using the LP-full protocol (regularisation selected on a held-out split, then refit on the full training set). ImageNet is validation accuracy. Natural shifts is the mean over ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b) and ObjectNet (Barbu et al., 2019) under both heads. ImageNet-O (Hendrycks et al., 2021b) is left out because its classes lie outside ImageNet-1k, so the probe has no outputs for them. ZS values use each shift's own label space and LP values the 1000-class label space, so levels can be compared across heads for ImageNet and ImageNet-C but not for natural shifts.

		ImageNet		ImageNet-C		Natural shifts	
Filter	Method	ZS	LP	ZS	LP	ZS	LP
Bestpool	Sequential	0.656	0.698	0.385	0.391	0.547	0.381
	Cumulative	0.708	0.721	0.454	0.436	0.603	0.413
	Oracle	0.718	0.728	0.470	0.453	0.625	0.430
Basic	Sequential	0.591	0.639	0.309	0.287	0.488	0.319
	Cumulative	0.623	0.643	0.346	0.303	0.524	0.327
	Oracle	0.638	0.653	0.349	0.311	0.535	0.336

Table 23: **Gains along the training ladder under both heads, with break-even slopes.** Δ is the change in accuracy (points) between consecutive checkpoints in Table 22. S $\rightarrow$ C adds replay. C $\rightarrow$ O retrains from scratch on all years, which also uses about four times the training compute. The break-even slope is the gain in the log-odds of shifted accuracy divided by the gain in the log-odds of ImageNet accuracy; a step raises effective robustness under any baseline whose slope in log-odds is below this value. For scale, a line through the six models has a slope of 1.2 (ZS) and 1.7 (LP) on ImageNet-C, and 0.9 (ZS) and 1.1 (LP) on natural shifts. Natural shifts use the mean accuracy over the five shifts in Table 22.

				ImageNet-C		Natural shifts	
Head	Step	Filter	ΔImageNet	Δ	Break-even	Δ	Break-even
ZS	S $\rightarrow$ C	Bestpool	+5.3	+7.0	1.2	+5.7	1.0
	S $\rightarrow$ C	Basic	+3.2	+3.7	1.3	+3.7	1.1
	C $\rightarrow$ O	Bestpool	+1.0	+1.5	1.3	+2.2	2.0
	C $\rightarrow$ O	Basic	+1.5	+0.3	0.2	+1.1	0.7
LP	S $\rightarrow$ C	Bestpool	+2.3	+4.5	1.7	+3.3	1.2
	S $\rightarrow$ C	Basic	+0.4	+1.7	5.3	+0.8	2.4
	C $\rightarrow$ O	Bestpool	+0.7	+1.7	2.0	+1.6	2.0
	C $\rightarrow$ O	Basic	+1.0	+0.7	0.8	+0.9	1.0

Table 24: k-shot CoOp accuracy (%), mean over three seeds. Within a seed, all six checkpoints use the same k-shot subset (paired design). *full*: all training images. The prompt comprises 16 context tokens initialized with "a photo of a", optimized with SGD (momentum 0.9, learning rate 2×10^{-3}, cosine schedule) for up to 200 epochs with early stopping on the validation split; the best-validation prompt is restored for test evaluation. $\bar{s}$: across-seed standard deviation averaged over the eleven datasets.

k	Model	IN	Cal	Pet	Car	Flo	Food	Air	SUN	DTD	ESAT	UCF	Mean	Med.	$\bar{s}$
Bestpool															
1	Seq.	65.5	94.5	87.5	80.1	79.1	80.9	22.5	63.8	54.5	59.2	59.4	67.9	65.5	1.16
	Cum.	70.6	95.6	89.6	87.2	82.8	87.2	27.1	68.9	58.4	65.0	65.5	72.5	70.6	1.08
	Ora.	72.2	96.1	90.7	87.3	87.7	88.8	31.1	69.3	60.3	47.4	69.9	72.8	72.2	1.99
4	Seq.	66.9	95.2	88.9	81.7	91.1	83.1	32.5	66.9	63.7	78.3	67.6	74.2	78.3	0.71
	Cum.	71.8	96.3	90.8	88.1	93.5	88.4	37.7	71.3	65.6	78.9	72.2	77.7	78.9	0.87
	Ora.	73.1	96.6	92.4	89.0	94.5	89.4	39.7	72.5	68.0	77.0	75.1	78.8	77.0	0.76
8	Seq.	68.1	95.9	89.7	84.2	93.9	84.2	37.1	69.0	66.7	83.9	72.1	76.8	83.9	0.57
	Cum.	72.3	96.6	91.3	89.0	96.2	89.1	41.3	72.8	70.6	83.8	75.5	79.9	83.8	0.60
	Ora.	73.7	96.9	92.1	89.2	96.7	89.7	39.8	74.2	72.0	83.7	78.0	80.6	83.7	0.69
16	Seq.	69.0	95.9	89.6	85.9	96.3	84.8	43.3	71.1	71.5	87.5	75.2	79.1	84.8	0.63
	Cum.	73.1	96.8	91.8	89.8	97.4	89.2	47.6	74.6	73.4	88.3	78.1	81.8	88.3	0.45
	Ora.	74.2	97.0	92.2	90.1	97.7	90.2	48.7	75.3	73.9	87.6	80.4	82.5	87.6	0.55
full	Seq.	72.3	96.3	90.7	86.4	96.0	88.0	43.9	73.0	72.6	95.2	79.7	81.3	86.4	0.00
	Cum.	75.8	97.0	92.1	88.9	97.0	91.4	49.2	75.7	74.7	95.4	80.7	83.4	88.9	0.00
	Ora.	76.6	97.0	92.8	91.3	97.3	91.7	48.6	76.5	77.8	95.3	83.9	84.4	91.3	0.00
Basic															
1	Seq.	59.1	91.9	82.4	78.2	67.3	77.6	7.9	62.1	47.5	57.9	57.0	62.6	62.1	1.45
	Cum.	62.5	93.5	82.8	83.9	71.4	81.9	15.7	65.3	52.0	50.4	58.6	65.3	65.3	1.44
	Ora.	64.2	94.3	84.3	83.7	74.9	83.2	12.6	66.2	56.0	51.7	61.1	66.6	66.2	2.47
4	Seq.	60.7	93.4	84.8	81.0	86.5	79.4	19.4	66.1	60.4	75.3	65.0	70.2	75.3	0.99
	Cum.	63.7	94.1	84.7	86.4	87.9	83.9	23.3	68.1	63.1	69.6	67.1	72.0	69.6	1.16
	Ora.	65.4	94.8	87.4	86.2	89.7	84.6	22.6	69.5	65.7	70.7	68.3	73.2	70.7	0.81
8	Seq.	61.8	93.6	85.8	83.5	91.0	81.2	22.0	68.2	64.4	79.8	69.9	72.8	79.8	0.69
	Cum.	64.4	94.4	85.7	86.8	92.2	84.3	27.0	70.2	68.9	77.3	70.9	74.7	77.3	0.82
	Ora.	66.4	94.8	87.8	87.3	93.8	85.3	27.2	71.2	68.0	78.1	69.9	75.4	78.1	0.70
16	Seq.	63.4	93.9	85.4	85.2	94.6	81.1	32.1	69.9	70.4	85.3	73.3	75.9	81.1	0.50
	Cum.	65.8	95.1	85.8	88.6	95.5	85.0	34.8	72.0	73.3	83.2	73.6	77.5	83.2	0.55
	Ora.	67.4	95.3	88.0	88.7	96.1	85.5	34.5	73.3	73.0	83.3	75.6	78.3	83.3	0.42
full	Seq.	67.1	94.5	87.5	86.0	94.2	86.8	30.1	71.9	74.6	93.7	76.8	78.5	86.0	0.00
	Cum.	69.4	95.0	87.7	88.9	94.4	88.2	33.2	72.3	76.1	93.0	77.3	79.6	87.7	0.00
	Ora.	70.8	95.4	89.3	89.0	96.8	89.5	30.8	74.5	76.0	94.6	77.9	80.4	89.0	0.00