GAINS ARE GAINS? WHAT DRIVES ACCURACY GAINS IN TRANSDUCTIVE CLIP METHODS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Transductive CLIP methods leverage unlabeled test batches to improve recognition, with performance gains typically summarized by aggregate classification accuracy. Yet the same accuracy gain can arise for very different reasons: a method may reduce errors caused by classes that are absent from the batch, improve discrimination among the classes that are present, or do both. We introduce a decomposition that separates these two effects and complement it with controlled interventions that vary the unlabeled context while holding key aspects of the prediction problem fixed. Across multiple fine-grained recognition datasets and CLIP variants, the resulting analysis reveals substantial differences that aggregate accuracy obscures. Methods with similar final performance can behave very differently. Moreover, a considerable fraction of the gain can remain even when the adapted scoring function is replaced by frozen CLIP scores while preserving the class set induced by the method. The controlled interventions further show that the balance between these effects depends strongly on the composition of the unlabeled context. Together, these findings show that final accuracy alone incompletely characterizes transductive adaptation.

1 INTRODUCTION

"In that case, I'll have the blueberry." The remark, commonly attributed to philosopher Sidney Morgenbesser, is a well-known illustration of the influence of irrelevant alternatives on decision making (Poundstone, 2008). The anecdote aims to show that the introduction of an apparently irrelevant option changes a decision that should otherwise remain unaffected. More generally, a decision can improve in two conceptually distinct ways: by removing irrelevant alternatives from consideration, or by improving the decision-making process among the relevant alternatives.

A closely related distinction arises in transductive image recognition. CLIP (Radford et al., 2021) enables zero-shot classification, where classes are predicted without task-specific labeled training examples, by comparing image representations with text-defined class representations. Transductive methods extend this setting by exploiting an unlabeled collection of test samples jointly, rather than treating every test image independently. Recent approaches such as EM-Dirichlet (Martin et al., 2024), TransCLIP (Zanella et al., 2024), StatA (Zanella et al., 2025), MOON (Huang et al., 2026b), and ADAPT (Zhang et al., 2025) use information available in the test batch to modify or refine predictions at inference time. Their improvements, however, are typically reported using global accuracy which does not reveal what actually drives the gains. Understanding this ambiguity is especially relevant when only a small subset of the full vocabulary (i.e., possible classes) is represented in a given test batch, because transductive methods can exploit batch-level information in ways that aggregate accuracy cannot distinguish. This observation motivates the central question of this work:

When a transductive CLIP method gains accuracy, what actually improved?

We address this question by decomposing transductive gains along two complementary axes. The first, which we refer to as *reduction in absent-class obstruction*, measures how much adaptation reduces errors caused by classes that are not represented in the current batch. We complement this quantity with predicted-support diagnostics that describe which classes appear in the method's outputs. These measurements characterize observable prediction behavior rather than the method's

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.

internal mechanism. The second, *within-support discrimination*, measures how well the adapted scoring function distinguishes among the classes that are actually represented in the batch.

To evaluate these capabilities, we complement the usual full-vocabulary accuracy with several metrics. First, after fitting every transductive method on the original full vocabulary, we evaluate its unchanged scores after restricting the candidate set to the classes represented in the batch. This evaluation-only true-support mask measures discrimination among represented classes without refitting or otherwise modifying the adapter. Second, we directly evaluate the *predicted support*, the set of classes appearing as top-1 predictions in the batch, using precision and recall. Finally, we introduce a *support-only* control that discards the adapted scoring function altogether: the predicted support is retained, but predictions are recomputed using unchanged frozen CLIP scores. Together, these metrics account for how a full-vocabulary gain relates to changes in within-support discrimination and to the reduction of errors caused by absent classes.

This decomposition reveals behavior that aggregate accuracy obscures. Methods with similar final accuracies can exhibit markedly different behavior along the two axes. More strikingly, a substantial fraction of the gain achieved by some transductive methods can subsist even when their adapted scoring functions are discarded. For example, on OpenAI ViT-B/32, MOON (Huang et al., 2026b) improves full-vocabulary accuracy from 61.14% to 64.17%, while frozen CLIP evaluated only over MOON's predicted support reaches 64.07%. At the same time, MOON's change in within-support accuracy is -0.93 percentage points, with an interval including zero. Thus, a substantial aggregate gain can occur without established evidence of improved discrimination among the classes represented in the batch. Under the MOON episodic protocol, where each test batch is treated as a separate episode, the distinction is sharper: with 64 images and one to four represented classes, its full-vocabulary accuracy gains 10.32 points while within-support accuracy falls by 3.09 points.

We further use the decomposition to study how transductive gains depend on the unlabeled context. Holding the evaluated query images fixed, we systematically change the surrounding test batch and observe substantial shifts in the balance between reduction in absent-class obstruction and within-support discrimination. These effects persist even when the set of classes represented in the batch is fixed and only the surrounding images change. Surprisingly, in some cases, context images from other classes that appear more similar to the target under frozen CLIP improve, rather than degrade, within-support discrimination. These results indicate that the effect of test-time context is strongly method-dependent and cannot be inferred from aggregate accuracy alone.

Our contributions are twofold:

- **Decomposing transductive gains.** We introduce an evaluation framework that distinguishes within-support discrimination from a reduction in absent-class obstruction. The framework combines full-vocabulary and within-support accuracy, predicted-support precision and recall, and a support-only control.
- **Revealing context-dependent behavior hidden by aggregate accuracy.** Using the proposed decomposition and controlled interventions, we show that transductive gains can change substantially with the composition of the unlabeled context, revealing that support and discrimination diagnostics can vary differently across methods and contexts in ways that final accuracy alone does not reveal.

2 Related work.

Recent transductive methods for CLIP differ in how they utilize an unlabeled test batch. EM-Dirichlet (Martin et al., 2024) models CLIP prediction vectors with class-specific Dirichlet distributions, while TransCLIP (Zanella et al., 2024) combines probabilistic modeling with graph-based transduction. StatA (Zanella et al., 2025) introduces a statistical anchor toward text-derived knowledge and studies settings with varying effective class counts and non-i.i.d. test streams. MOON (Huang et al., 2026b) uses a von Mises-Fisher mixture, a mixture model for normalized feature vectors, with dynamic shrinkage, including evaluations under class imbalance and batches containing only a subset of classes, while ADAPT (Zhang et al., 2025) performs backpropagation-free probabilistic Gaussian alignment. These differences suggest that transductive gains may arise for different reasons across methods. Methods such as StatA and MOON may benefit more from identifying which classes are supported by the batch, while methods such as ADAPT may rely more on improv-

ing discrimination among represented classes. These methods form the main transductive baselines in our study. Test-time adaptation for CLIP also includes methods based on different sources of evidence, such as TPT (Shu et al., 2022), which adapts prompts from augmented views of an individual image, and TDA (Karmanov et al., 2024), which accumulates test features and pseudo-labels in a dynamic cache. Our setting instead focuses on cross-sample information from a jointly observed unlabeled batch. Recent empirical work has investigated what drives CLIP adaptation across different update mechanisms and sources of test-time evidence (Huang et al., 2026a), while other analyses study properties such as calibration (Choi et al., 2026). Our question is complementary, when a transductive method improves recognition accuracy, what actually improved?

3 Decomposing Transductive Gains

Let $\mathcal{C}$ denote the full vocabulary, i.e., the full set of class labels available at inference time. Given an unlabeled test batch $B = \{x_i\}_{i=1}^N$ of N images, only a subset of possible classes might be actually represented in the batch. We refer to this subset as the *true support* and denote it by $S = \{y_i : x_i \in B\} \subseteq \mathcal{C}$. We write $K = |S|$ for the number of classes represented in the batch. A frozen CLIP model (Radford et al., 2021) denoted 0 provides a zero-shot score $s_0(x_i, c)$ for each image x_i and class $c \in \mathcal{C}$. Its prediction is obtained by selecting the highest-scoring class over the full vocabulary $\mathcal{C}$.

Rather than processing each $x_i \in B$ independently, transductive CLIP methods exploit the unlabeled batch jointly. Given B, a method M produces adapted scores $s_M(x_i, c; B)$ for $c \in \mathcal{C}$, so that the score assigned to an image depends on the other unlabeled batch images. The goal of these methods is to exploit information shared across the batch to improve recognition.

3.1 Full-Vocabulary and Within-Support Accuracy

The usual measure of transductive performance is recognition accuracy over the full vocabulary $\mathcal{C}$. For a fitted method M, we write

$$A_{\text{full}}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\max_{c \in \mathcal{C}} s_M(x_i, c; B) = y_i\right]. \tag{1}$$

To isolate discrimination among represented classes, we evaluate the same adapted scores after restricting prediction to the true support S:

$$A_{\text{supp}}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\max_{c \in S} s_M(x_i, c; B) = y_i\right]. \tag{2}$$

The method is fitted using the full vocabulary $\mathcal{C}$; the restriction to S is solely applied for evaluation purposes. Thus, $A_{\text{supp}}(M)$ measures discrimination among represented classes without providing the true support to the transductive method. Both quantities can also be evaluated on a designated query subset $Q \subseteq B$, which we use in the controlled context experiments. We also define

$$\Delta_{\text{full}}(M) = A_{\text{full}}(M) - A_{\text{full}}(0), \qquad \Delta_{\text{rank}}(M) = A_{\text{supp}}(M) - A_{\text{supp}}(0). \tag{3}$$

Here, Δ_{rank} measures the change in discrimination among classes represented in the batch.

The gap between within-support and full-vocabulary accuracy captures *absent-class obstruction*: cases in which some absent class $c \in \mathcal{C} \setminus S$ receives a higher score than the correct class, even though the correct class is top-ranked within the true support. We define :

$$R(M) = A_{\text{supp}}(M) - A_{\text{full}}(M). \tag{4}$$

Since restricting prediction to the true support can only preserve or repair a prediction, $R(M) \geq 0$. These quantities yield

$$\Delta_{\text{full}}(M) = \Delta_{\text{rank}}(M) + \underbrace{R(0) - R(M)}_{\Delta_R = \text{reduction in absent-class obstruction}}. \tag{5}$$

Thus, a full-vocabulary gain can reflect improved discrimination among represented classes, reduced obstruction from absent classes, or both. Equation 5 is an exact algebraic identity obtained by direct substitution of the definitions above; Appendix A details the image-level derivation.

3.2 Measuring Predicted Support Directly

The decomposition describes how absent classes affect recognition outcomes, but does not directly measure which classes the method identifies as represented. We therefore formalize the *predicted support* as:

$$\widehat{S}_M = \left\{ \arg\max_{c \in \mathcal{C}} s_M(x_i, c; B) \ : \ x_i \in B \right\}. \tag{6}$$

We compare $\widehat{S}_M$ with the true support using precision and recall:

$$P_S(M) = \frac{|\widehat{S}_M \cap S|}{|\widehat{S}_M|}, \qquad R_S(M) = \frac{|\widehat{S}_M \cap S|}{|S|}. \tag{7}$$

Precision measures how many predicted classes are represented in the batch, while recall measures how many represented classes are recovered. Since $\widehat{S}_M$ is induced from the method's hard predictions, it should be interpreted as an observable summary of its output rather than an explicit latent support estimate.

3.3 Support-Only Control

Support precision and recall measure how closely the predicted support matches the true support, but not how useful that class set is for recognition. We therefore introduce a *support-only* control that retains $\widehat{S}_M$ while discarding the adapted scoring function:

$$\widehat{y}_i^{\mathrm{SO}}(M) = \arg\max_{c \in \widehat{S}_M} s_0(x_i, c), \qquad A_{\mathrm{SO}}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\widehat{y}_i^{\mathrm{SO}}(M) = y_i\right]. \tag{8}$$

As with the previous accuracy measures, A_{SO} can be evaluated on a designated query subset $Q \subseteq B$. Unlike within-support accuracy, this control does not use the true support: the transductive method contributes only $\widehat{S}_M$, while all image–class scores are supplied by frozen CLIP. If $A_{\mathrm{SO}}(M)$ approaches $A_{\mathrm{full}}(M)$, then much of the method's predictive value is already captured by the induced class set. A larger gap indicates additional contribution from the adapted scoring function.

4 Controlled Evaluation Design

Our experiments follow the two terms of the decomposition in Equation 5. We first test whether separating within-support discrimination from absent-class obstruction reveals behavior hidden by full-vocabulary accuracy in an established transductive setting. We then intervene on the unlabeled context around fixed query images, first by changing which classes are represented in a batch and then by fixing those classes and changing only their images. Finally, we vary support sparsity, i.e., how few classes are represented in the batch, and track both terms of the decomposition as the number of represented classes changes.

We evaluate CUB-200-2011 (Wah et al., 2011), FGVC-Aircraft (Maji et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), Flowers-102 (Nilsback & Zisserman, 2008), and Food-101 (Bossard et al., 2014) using OpenAI CLIP ViT-B/32 (Radford et al., 2021). Results for OpenAI CLIP ViT-B/16 and LAION-2B ViT-B/32 (Schuhmann et al., 2022) are reported in Appendix Sections C–E. Comparators are frozen CLIP, a pseudo-label linear head, ZLaP (Stojnić et al., 2024), GDA-CLIP (Wang et al., 2024), TransCLIP (Zanella et al., 2024) , ADAPT (Zhang et al., 2025), StatA (Zanella et al., 2025), MOON (Huang et al., 2026b), and MOON (soft). Within each protocol and checkpoint, methods share the same frozen representations and text head. We begin with the MOON episodic protocol, using 1,000 shared episodes per dataset, where each episode is one sampled unlabeled test batch with its associated class subset. Methods receive only the unlabeled batch and full vocabulary and are reset between episodes, more details are provided in Appendix B. This setting lets us evaluate the full set of proposed diagnostics on the same predictions, including full-vocabulary accuracy, within-support accuracy, absent-class obstruction, predicted-support precision and recall, and support-only accuracy.

Context intervention. To study semantic context, every class in each dataset serves once as the target class, giving 540 target-class episodes per context across the five datasets. For each target class, we select eight query images once and reuse the same queries in the near, random, and far contexts. Each query set is surrounded by eight images from each of seven other classes ($N = 64$, $K = 8$), chosen to be near, random, or far from the target in frozen OpenAI ViT-B/32 text-embedding space. Thus, across the three contexts, the evaluated queries, batch size, class counts, and full vocabulary are fixed; only the seven surrounding classes and their images change. We then use a stricter fixed-support intervention. For each target, we keep the query images and the same seven surrounding class identities fixed. Within each surrounding class, however, we replace its context images with examples that frozen OpenAI CLIP assigns low, random, or high probability to the target class. The true support therefore stays unchanged; only the images representing the surrounding classes vary. This lets us test whether those images alone can change discrimination.

Support sparsity. We vary support sparsity using MOON episodic regimes: $N = 64$ with $K \in \{1\text{–}4, 2\text{–}10, 5\text{–}25\}$ and $N = 1{,}000$ with $K \in \{5\text{–}25, 25\text{–}50, 50\text{–}100\}$. This lets us measure how both Δ_{rank} and Δ_R change as the true support broadens.

We average episode-level metrics within each dataset and then average with equal weights across datasets. We report pointwise 95% confidence intervals from 10,000 bootstrap resamples of episodes within each dataset and target classes for the context interventions. Details in Appendix B

5 Results

5.1 The same accuracy gain can reflect different improvements

We begin with two contrasting regimes from the established transductive setting of MOON Huang et al. (2026b). In Table 1, the sparse regime ($N = 64$, $K = 1\text{–}4$), MOON improves full-vocabulary accuracy from 61.72% to 72.04%. Yet its within-support accuracy falls by 3.09 points. Equation 5 accounts for the full-vocabulary gain exactly: $+10.32$ percentage points. The positive full-vocabulary gain is therefore accompanied by worse discrimination among represented classes and a much larger reduction in absent-class obstruction. Frozen CLIP predictions restricted only to MOON's predicted support in the SO column reach 70.89%, retaining most of adapted MOON's results without using the scores, further showing that in this regime, much of MOON's predictive benefit is already captured by its predicted support rather than by improved discrimination among the classes in the true support.

In Table 2, the broad-support regime ($N = 1{,}000$, $K = 50\text{–}100$) shows a different pattern. ADAPT improves within-support accuracy by $+1.84$ points, but absent-class obstruction worsens by 0.93 points, leaving a full-vocabulary gain of only $+0.91$ points. Most strikingly, MOON and ADAPT reach 63.60% and 62.85% full-vocabulary accuracy, respectively, in this regime, but their decompositions point in opposite directions: MOON has $\Delta_{\text{rank}} = -1.45$ and $\Delta_R = +3.12$, whereas ADAPT has $\Delta_{\text{rank}} = +1.84$ and $\Delta_R = -0.93$, showing that similar full vocabulary gains hide distinct behaviors. GDA-CLIP and TransCLIP show similar a behavior to ADAPT, with positive Δ_{rank} but negative Δ_R, whereas StatA and MOON (soft) exhibit a pattern similar to MOON.

Table 1: MOON-style evaluation with sparse support ($N = 64$, $K = 1\text{–}4$), OpenAI ViT-B/32. Each regime uses 1,000 shared episodes per dataset across five datasets (5,000 episodes). Entries are mean [95% CI]. Full and support-only (SO) are percentages; Δ_{rank} and Δ_R are percentage points.

Method	Full		Δ_{rank}		Δ_R		SO	
Frozen CLIP	61.72	[61.21, 62.24]	+0.00	[+0.00, +0.00]	+0.00	[+0.00, +0.00]	61.72	[61.21, 62.24]
Pseudo-linear	62.25	[61.72, 62.78]	−1.26	[−1.37, −1.14]	+1.79	[+1.65, +1.92]	61.97	[61.44, 62.49]
ZLaP	22.45	[22.15, 22.74]	−1.18	[−1.36, −1.01]	−38.10	[−38.61, −37.59]	60.22	[59.61, 60.82]
GDA-CLIP	31.36	[31.06, 31.66]	−4.60	[−4.77, −4.44]	−25.76	[−26.15, −25.38]	62.64	[62.10, 63.17]
TransCLIP	36.20	[35.87, 36.53]	−5.52	[−5.86, −5.17]	−20.35	[−20.86, −19.84]	62.07	[61.56, 62.58]
ADAPT	56.84	[56.34, 57.33]	−4.41	[−4.74, −4.09]	−0.47	[−0.85, −0.09]	61.72	[61.21, 62.24]
StatA	69.32	[68.71, 69.92]	−1.11	[−1.24, −0.99]	+8.71	[+8.40, +9.02]	67.42	[66.83, 68.00]
MOON	72.04	[71.43, 72.66]	−3.09	[−3.31, −2.87]	+13.41	[+13.06, +13.76]	70.89	[70.28, 71.51]
MOON (soft)	72.47	[71.86, 73.07]	−3.05	[−3.27, −2.83]	+13.79	[+13.45, +14.13]	71.50	[70.90, 72.10]

Table 2: MOON-style evaluation with broader support ($N = 1{,}000$, $K = 50$–100), OpenAI ViT-B/32. Metrics and units match Table 1

Method	Full		Δ_{rank}		Δ_R		SO	
Frozen CLIP	61.94	[61.87, 62.00]	+0.00	[+0.00, +0.00]	+0.00	[+0.00, +0.00]	61.94	[61.87, 62.00]
Pseudo-linear	62.36	[62.29, 62.43]	−0.50	[−0.52, −0.48]	+0.92	[+0.90, +0.95]	62.03	[61.96, 62.10]
ZLaP	58.03	[57.95, 58.11]	−0.16	[−0.20, −0.13]	−3.74	[−3.81, −3.68]	62.22	[62.16, 62.29]
GDA-CLIP	60.87	[60.80, 60.94]	+1.80	[+1.77, +1.84]	−2.87	[−2.92, −2.82]	61.95	[61.88, 62.01]
TransCLIP	62.45	[62.39, 62.52]	+1.53	[+1.49, +1.56]	−1.01	[−1.05, −0.98]	61.94	[61.87, 62.00]
ADAPT	62.85	[62.78, 62.92]	+1.84	[+1.81, +1.86]	−0.93	[−0.95, −0.90]	61.94	[61.87, 62.00]
StatA	63.91	[63.83, 63.99]	−0.67	[−0.71, −0.63]	+2.64	[+2.60, +2.69]	62.48	[62.40, 62.55]
MOON	63.60	[63.52, 63.68]	−1.45	[−1.49, −1.42]	+3.12	[+3.07, +3.16]	62.77	[62.69, 62.84]
MOON (soft)	63.11	[63.02, 63.19]	−1.79	[−1.82, −1.76]	+2.96	[+2.92, +3.00]	62.47	[62.39, 62.54]

Finding 1: A positive full-vocabulary gain does not identify what improved. The same gain can reflect better within-support discrimination, reduced absent-class obstruction, or opposing changes in the two. In some regimes, most of the gain is retained by the predicted support even after the adapted scores are discarded.

5.2 Context Interventions: The effect of the context in the batch on transductive gains

We next ask how the unlabeled context of the same query images changes transductive performance. We study this at two levels, as detailed in Section 4: (1) by changing which classes are represented in the batch, and (2) by fixing those classes and changing only the images that represent them.

5.2.1 Transductive gains under fine-grained semantic confusability

We first examine whether the semantic similarity of the represented classes changes both the magnitude and the decomposition of transductive gains. In order to isolate the effect of semantically confusable classes, we use the context intervention explained in Section 4 to hold the evaluated query images fixed while changing the other classes represented in the unlabeled batch. The evaluated query images, batch size, class counts, and full vocabulary are fixed. Only the seven surrounding classes and their 56 context images change.

This intervention changes how difficult it is to discriminate among the represented classes without changing frozen full-vocabulary predictions. On OpenAI ViT-B/32, frozen full-vocabulary accuracy is therefore 61.14% in all three contexts. However after restricting predictions to the true support and calculating A_{supp}, frozen CLIP reaches only 67.69% with near classes, compared with 87.16% with random classes and 97.24% with far classes (Figure 1), confirming that with near classes, the within-support discrimination becomes substantially more difficult. This shift also changes where the remaining errors lie. In near context, frozen CLIP has only 6.55 points of absent-class obstruction. This grows to 26.02 points under random context and 36.10 points under far context. When the represented classes are difficult to distinguish, much of the remaining error persists even after absent classes are removed. When within-support discrimination is already easy, absent alternatives account for a much larger part of the gap to full-vocabulary accuracy.

Table 3 examines the difficult near context. StatA and MOON obtain positive full-vocabulary gains while their changes in within-support accuracy are negative. For MOON, the gain is +3.03 points, accompanied by a reduction in absent-class obstruction of +3.96 points. When the represented alternatives are deliberately made difficult, almost all methods start to discriminate less accurately among the classes that are actually present. Notably, TransCLIP achieves very high predicted-support recall (97.36%) despite negative Δ_{rank} and Δ_R; its low support precision (21.08%) indicates that this recall comes with many absent classes retained in the predicted support.

Table 3: Near-context queries, averaged equally over five datasets. OpenAI ViT-B/32 uses 540 target-class episodes (540 episode runs). Entries are mean [95% CI]. Full and SO are percentages; Δ_{rank} and Δ_R are percentage-point changes relative to frozen CLIP and are evaluated on Q, the eight fixed target-class queries in each $N = 64$ episode. P and R are predicted-support precision and recall.

Method	Full	Δ_{rank}	Δ_R	SO	P	R
Frozen CLIP	61.14 [58.68, 63.56]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.14 [58.68, 63.56]	48.92	89.10
ADAPT	60.24 [57.69, 62.65]	−1.47 [−2.61, −0.34]	+0.57 [−0.25, +1.36]	61.08 [58.56, 63.53]	48.83	86.21
StatA	61.67 [59.00, 64.25]	−2.04 [−3.76, −0.39]	+2.57 [+1.57, +3.60]	62.85 [60.32, 65.30]	63.75	81.41
MOON	64.17 [61.38, 66.88]	−0.93 [−2.13, +0.30]	+3.96 [+2.97, +4.96]	64.07 [61.38, 66.66]	72.32	73.41
MOON (soft)	63.95 [61.17, 66.66]	−1.84 [−2.99, −0.72]	+4.65 [+3.66, +5.70]	64.17 [61.43, 66.77]	77.66	73.97
TransCLIP	44.54 [42.21, 46.85]	−1.80 [−3.49, −0.08]	−14.80 [−16.46, −13.21]	61.14 [58.68, 63.56]	21.08	97.36

Interestingly, the sparse regime in Table 1 can hide this difficulty. On the same OpenAI ViT-B/32 checkpoint, frozen CLIP already reaches 96.01% within-support for $N = 64$, $K = 1$–4 (Appendix Table 5). MOON gains $+10.32$ points in full vocabulary in that setting, with $\Delta_{\text{rank}} = -3.09$ and $\Delta_R = +13.41$, so the large full-vocabulary gain appears almost entirely through reduced absent-class obstruction. In near context, by contrast, frozen within-support accuracy falls to 67.69%, and MOON's gain contracts to $+3.03 = -0.93 + 3.96$ points. Thus, a sparse evaluation in which the represented classes are already easy to distinguish can obscure how a method behaves when the true alternatives themselves are semantically confusable.

We next use the full near/random/far intervention to examine how methods respond as within-support discrimination becomes progressively easier. Figure 1 reports the two terms of the decomposition for each method.

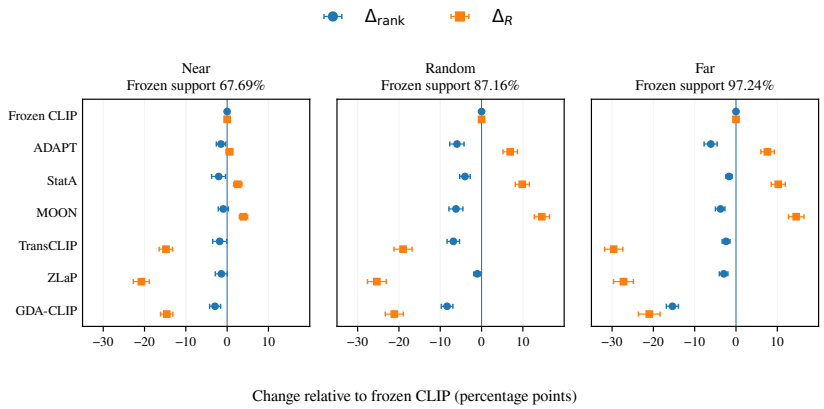

Figure 1: OpenAI ViT-B/32 context intervention with identical query images across near, random, and far contexts. Circles show the change in within-support accuracy, Δ_{rank}, and squares show the reduction in absent-class obstruction, Δ_R, both relative to frozen CLIP. Frozen within-support accuracy is 67.69, 87.16, and 97.24 in the near, random, and far contexts, respectively. Horizontal bars show 95% bootstrap intervals.

We notice two behaviors, for StatA and MOON, the absolute reduction in absent-class obstruction becomes substantially larger as the represented classes become easier to distinguish. MOON's Δ_R, for example, increases from $+3.96$ points in near context to $+14.58$ under random context and $+14.57$ under far context. StatA increases from $+2.57$ to $+9.85$ and $+10.23$ points. ADAPT shows the same direction at smaller scale, with Δ_R increasing from $+0.57$ to $+6.94$ and $+7.61$ points. These larger absolute reductions in absent-class obstruction occur while Δ_{rank} remains negative. Because the baseline amount of absent-class obstruction also differs across contexts, read these values as absolute contributions to accuracy rather than normalized removal efficiencies. TransCLIP, ZLaP, and GDA-CLIP show the opposite response. Their absent-class obstruction increases rather than decreases, and the effect becomes larger as the represented classes become easier to distinguish.

Finding 2: Semantic confusability changes both the size and the source of transductive gains. StatA and MOON show larger absolute reductions in absent-class obstruction as the represented classes become easier to distinguish, while TransCLIP, ZLaP, and GDA-CLIP show the opposite pattern.

5.2.2 More confusable context images can improve discrimination

To isolate the effect of the context images from the identities of the represented classes, we fix the query images and all eight class identities from the near context and vary only the images drawn from the seven surrounding classes. Within each surrounding class, we select images with low, random, or high frozen-CLIP probability for the target class. The query images, true support, and frozen query predictions are therefore unchanged across the three contexts.

Table 4 shows that changing only the context images can substantially alter within-support discrimination. For MOON, Δ_{rank} moves from -3.06 with low-target-probability context to $+3.84$ with high-target-probability context, a $+6.90$-point shift. StatA shows the same qualitative reversal, from -0.54 to $+2.87$. Thus, for both methods, adaptation worsens within-support discrimination when the surrounding examples are relatively unlike the target, but becomes beneficial when those examples are more target-like under frozen CLIP. ADAPT does not show this behavior, with Δ_{rank} remaining negative across all three contexts, the effect is therefore method dependent.

Table 4: Effect of context-image selection on within-support discrimination, OpenAI ViT-B/32. Query images and represented classes are fixed; only the context images change. Entries are Δ_{rank} in percentage points relative to frozen CLIP. There are 540 paired target classes across five datasets (1,620 episode runs across three conditions). Intervals are 95% percentile CIs.

Method	Low		Random		High		High–Low	
StatA	-0.54	$[-2.12, +1.07]$	-2.11	$[-3.66, -0.57]$	$+2.87$	$[+1.16, +4.60]$	$+3.41$	$[+1.68, +5.26]$
MOON	-3.06	$[-4.45, -1.78]$	-2.08	$[-3.47, -0.85]$	$+3.84$	$[+2.50, +5.21]$	$+6.90$	$[+5.28, +8.60]$
ADAPT	-0.86	$[-1.82, +0.11]$	-1.37	$[-2.49, -0.22]$	-1.25	$[-2.46, -0.05]$	-0.39	$[-1.53, +0.74]$

Finding 3: More confusable context can help rather than hurt. For MOON and StatA, context images from competing classes that look more like the target improve discrimination among the same represented classes.

5.3 The effect of support sparsity on transductive gains

To test how support sparsity affects transductive gains, we fix the batch size at $N = 64$ and vary the number of represented classes as $K \in \{4, 8, 16, 32\}$. For each K, we evaluate 1,000 shared episodes per dataset on official test images and evaluate every image in the batch. Figure 2 shows a strong dependence on support sparsity for MOON and StatA. On OpenAI ViT-B/32, MOON's full-vocabulary gain falls from $+10.38$ points at $K = 4$ to -2.15 at $K = 32$, while StatA falls from $+7.17$ to -1.87. ADAPT remains comparatively stable, moving from 1.43 to $+0.02$ points.

The decomposition explains why the sparse setting is particularly favorable for MOON and StatA. For MOON, Δ_R is $+15.49$ points at $K = 4$ but falls to $+7.40$ at $K = 32$, while Δ_{rank} is negative throughout and decreases from -5.10 to -9.55. StatA shows the same structure: Δ_R falls from $+9.16$ to $+3.68$, while its within-support change remains negative. Thus, when few classes are represented, there is substantially more room for these methods to gain by reducing errors due to absent classes. As the support broadens, that contribution shrinks and no longer compensates for the loss in within-support discrimination. ADAPT's two components remain much more stable across K, showing that exploiting this sparse-support opportunity is method-dependent.

Finding 4: Sparse support creates more room to reduce absent-class obstruction. For MOON and StatA, Δ_R is largest when few classes are represented and shrinks as the support broadens, while within-support discrimination remains worse than frozen CLIP. This large Δ_R more than offsets Δ_{rank} loss, so the degradation in discrimination is masked by a large full-vocabulary accuracy gain. ADAPT remains comparatively stable across K.

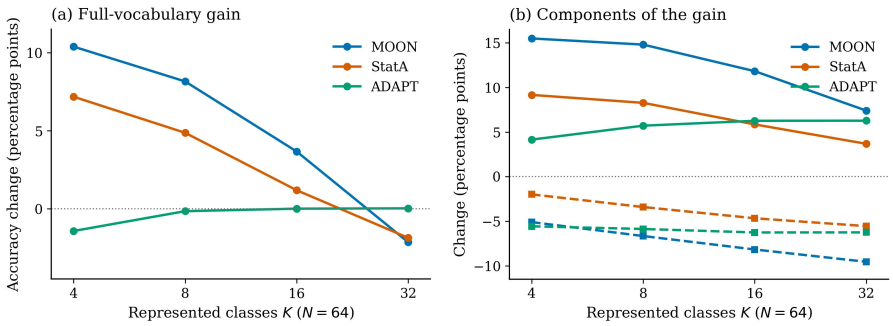

Figure 2: Fixed-N support-sparsity intervention on OpenAI ViT-B/32 with balanced class counts, averaged over five datasets. Each condition uses $N = 64$ and 1,000 shared episodes per dataset. (a) Full-vocabulary accuracy change relative to frozen CLIP. (b) Reduction in absent-class obstruction, Δ_R (solid), and change in within-support accuracy, Δ_{rank} (dashed). Intervals for both components appear in Appendix Table 61; the $N = 32$ and $N = 96$ results in Appendix Tables 65 and 69 show the same pattern.

6 Discussion

Our results suggest that transductive CLIP methods should be viewed as providing different capabilities rather than as interchangeable ways of improving aggregate accuracy. Some methods gain primarily by reducing errors from classes that are absent from the batch, whereas others are more effective at improving discrimination among represented classes. This distinction matters because the balance between the two effects changes with support sparsity and semantic context. In deployments where the expected batch structure is approximately known, the decomposition can therefore help characterize which methods are better matched to that setting. Methods that gain mainly through reduced absent-class obstruction may be particularly useful when only a small fraction of the vocabulary is expected to appear, whereas within-support discrimination becomes more important as the represented support broadens. We therefore recommend reporting within-support accuracy and absent-class obstruction alongside full-vocabulary accuracy when evaluating transductive methods.

Several limitations qualify these conclusions. The proposed measurements are diagnostic rather than mechanistic: they identify where performance changes occur, but not why an adaptation method produces them. Our experiments cover five recognition datasets and multiple CLIP checkpoints, including OpenAI ViT-B/16, OpenAI ViT-B/32, and LAION ViT-B/32, but remain within the ViT-B CLIP family. It therefore remains to be seen how broadly the observed patterns extend to larger architectures, other vision–language model families, and different domains. Finally, our controlled episodic protocols manipulate support sparsity and context composition, whereas real deployment streams may contain more complex and evolving class distributions.

7 Conclusion

Full-vocabulary accuracy alone provides an incomplete picture of transductive CLIP adaptation. We introduced an exact decomposition that separates changes in within-support discrimination from changes in absent-class obstruction, and used it to reveal behaviors that aggregate accuracy obscures. Across methods, support regimes, and controlled context interventions, similar accuracy gains can arise from very different changes in the prediction problem, and the balance between these effects can shift substantially with the composition of the unlabeled batch. These results motivate evaluating transductive methods not only by how much accuracy they gain, but also by where those gains come from.

AI USE STATEMENT

Generative AI tools assisted with research ideation, experiment design, code development and debugging, literature discovery, and drafting and editing of the manuscript and tables. All claims in this draft were checked against the saved experiment analyses. The authors are responsible for verifying the experimental procedures, sources, and final text.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. URL https://data.vision.ee.ethz.ch/cvl/vision2/datasets_extra/food-101/static/bossard_eccv14_food-101.pdf.

Jungwon Choi, Hyeonseo Jang, Kibok Lee, and Eunwoo Kim. Rethinking the test-time prompt tuning objective from the perspective of calibration. *arXiv preprint arXiv:2608.30230*, 2026. arXiv:2608.30230v2.

Jiazhen Huang, Xiao Chen, Zhiming Liu, Yaru Sun, Jingyan Jiang, and Zhi Wang. What drives test-time adaptation for CLIP? a controlled empirical study from an update perspective. *arXiv preprint arXiv:2606.14299*, 2026a. URL https://arxiv.org/abs/2606.14299.

Jiazhen Huang, Zhiming Liu, Changhu Wang, Wei Ju, Ziyue Qiao, and Xiao Luo. Von mises-fisher mixture model with dynamic shrinkage for realistic test-time transduction. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Karmanov_Efficient_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2024_paper.html.

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

Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2405.18437.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. URL https://www.robots.ox.ac.uk/~vgg/publications/2008/Nilsback08/nilsback08.pdf.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/parkhi12a.pdf.

William Poundstone. *Gaming the Vote: Why Elections Aren't Fair (and What We Can Do About It)*. Hill and Wang, 2008.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev.

LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/abs/2210.08402.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.07511.

Vladan Stojnić, Yannis Kalantidis, and Giorgos Tolias. Label propagation for zero-shot classification with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23209–23218, 2024. doi: 10.1109/CVPR52733.2024.02190. URL https://arxiv.org/abs/2404.04072.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.04087.

Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62223–62256, 2024. URL https://arxiv.org/abs/2406.01837.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zanella_Realistic_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2025_paper.html.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2562.

A AN EXACT DECOMPOSITION OF RECOGNITION ACCURACY

We derive Equation 5 for an arbitrary evaluated set $E \subseteq B$; in the episodic experiments $E = B$, whereas in the matched-query experiments $E = Q$. Let $\hat{y}^{\mathcal{C}}_{M,i}$ and $\hat{y}^{S}_{M,i}$ denote argmax predictions from the same fitted scores over the full vocabulary and the true support, respectively. Since $y_i \in S \subseteq \mathcal{C}$, $\hat{y}^{\mathcal{C}}_{M,i} = y_i$ implies $\hat{y}^{S}_{M,i} = y_i$. Hence the following identity holds for each evaluated image:

$$\mathbf{1}[\hat{y}^{\mathcal{C}}_{M,i} = y_i] = \mathbf{1}[\hat{y}^{S}_{M,i} = y_i] - \mathbf{1}[\hat{y}^{S}_{M,i} = y_i,\, \hat{y}^{\mathcal{C}}_{M,i} \neq y_i]. \tag{9}$$

The second indicator identifies an image classified correctly when absent classes are excluded, but incorrectly when they are admitted. Averaging over E gives

$$A^{E}_{\text{full}}(M) = A^{E}_{\text{support}}(M) - R^{E}(M), \qquad R^{E}(M) = \frac{1}{|E|} \sum_{i \in E} \mathbf{1}[\hat{y}^{S}_{M,i} = y_i,\, \hat{y}^{\mathcal{C}}_{M,i} \neq y_i] \geq 0. \tag{10}$$

Subtracting this equality for frozen CLIP (0) from the equality for M yields

$$A^{E}_{\text{full}}(M) - A^{E}_{\text{full}}(0) = \left[A^{E}_{\text{support}}(M) - A^{E}_{\text{support}}(0)\right] + \left[R^{E}(0) - R^{E}(M)\right] \tag{11}$$

$$= \Delta^{E}_{\text{rank}}(M) + \Delta^{E}_{R}(M). \tag{12}$$

The identity is exact per episode and therefore remains exact after averaging episodes or datasets with the same weights for all three terms. The terms are signed: a method can improve within-support ranking while increasing absent-class obstruction, or conversely. Although $R^{E}(M)$ is non-negative, $\Delta^{E}_{R}(M)$ need not be. The support-only accuracy A_{SO} is a separate counterfactual that replaces the adapted scores with frozen CLIP scores while retaining $\widehat{S}_M$; it is neither a term of Equation 12 nor an upper bound on adapted accuracy.

B IMPLEMENTATION DETAILS

B.1 DATASETS, CHECKPOINTS, AND TEXT REPRESENTATIONS

We use the official test splits of CUB-200-2011 (5,794 images and 200 classes), FGVC-Aircraft (3,333 images and 100 classes), Oxford-IIIT Pets (3,669 images and 37 classes), Flowers-102 (6,149 images and 102 classes), and Food-101 (25,250 images and 101 classes).

The primary checkpoint is OpenAI CLIP ViT-B/32. We additionally evaluate OpenAI CLIP ViT-B/16 and LAION-2B ViT-B/32 . Image embeddings are extracted once for each checkpoint, stored in single precision, and normalized to unit length. Methods compared within a condition receive identical image embeddings and text classifiers. The native MOON protocol uses the released dataset-specific template. The controlled context and fixed-N experiments use the normalized mean of 16 generic CLIP prompt embeddings for each class. The encoders remain frozen throughout.

Every adapter receives the complete class vocabulary and the unlabeled images in the current episode. It does not receive labels, the true support, or the query mask. Labels are used only to construct controlled episodes and to evaluate predictions after adaptation. Each episode begins with a fresh method state.

B.2 PUBLISHED IMPLEMENTATIONS AND COMPATIBILITY

We obtained the native sampler and common comparison solvers from the authors' official MOON repository[1]. This release contains the implementations used by MOON to compare ZLaP, GDA-CLIP, TransCLIP, ADAPT, and StatA under a common interface. This source avoids us reimplementing the published objectives from paper descriptions alone and make the evaluated code path explicit.

Compatibility changes are restricted to execution environment and numerical linear algebra. Hard-coded device transfers are redirected to the selected device. ZLaP uses the released graph equations with PyTorch inner-product nearest neighbors and a SciPy sparse direct solve in place of

[1] https://github.com/walawalagoose/MOON/

FAISS/CuPy. TransCLIP is replayed in double precision when its single-precision output is non-finite. These changes leave the score definitions and method inputs unchanged. The pseudo-linear control is our own frozen-feature baseline: it is initialized from the text classifier, fitted to hard frozen-CLIP pseudo-labels for 40 steps with learning rate 0.001, and regularized toward its initial weights.

B.3 HYPERPARAMETERS

The native experiments use the released comparison settings. ZLaP uses $k = 5$, $\gamma = 5$, and $\alpha = 0.3$; GDA-CLIP uses $\alpha = 5$; TransCLIP uses $\lambda_{\hat{y}} = 1$, $\lambda_{\text{Laplacian}} = 1$, and three neighbors; ADAPT uses $\alpha = 0.9$ and bank size 12. StatA and MOON use $\alpha = 1$, $\lambda_{\hat{y}} = 1$, $\lambda_{\text{Laplacian}} = 1$, three neighbors, and ten update iterations. MOON (soft) changes only the released `soft_beta` option.

The context interventions use settings selected before the reported experiments. We evaluated a common grid on 108 development episodes from CUB, Aircraft, and Pets, spanning approximately i.i.d. and concentrated-class batches. A single setting per method was selected from 12 candidates and then held fixed: StatA uses $\alpha = 0.1$ and graph weight 5; MOON uses $\alpha = 1$ and graph weight 0; ADAPT uses $\alpha = 0.1$, scale 200, and bank size 6. MOON (soft) again changes only `soft_beta`. ZLaP, GDA-CLIP, and TransCLIP retain the released settings above and were not tuned on the context intervention. No method is selected or tuned using the test episodes reported below.

B.4 EPISODE CONSTRUCTION AND STATISTICAL ANALYSIS

The native experiment follows the released MOON sampler. For each of six (N, K) regimes, it draws 1,000 episodes per dataset from the official test split. The same episode manifests are shared by all methods and checkpoints, and every image in an episode is evaluated. The primary statistic is the mean episode metric within each dataset followed by an equally weighted mean of the five dataset values.

The context intervention uses every class once as the target, giving 540 target-class episodes per condition across the five datasets. Each episode contains eight fixed target queries and 56 context images. The target queries are identical across near, random, and far conditions; the seven context classes are chosen by their frozen OpenAI ViT-B/32 text-embedding similarity to the target. The stricter fixed-class intervention retains these class identities and replaces only the context images, selecting examples with low, random, or high frozen-CLIP probability for the target class.

The fixed-N support-sparsity experiment uses 1,000 shared official-test episodes per dataset, checkpoint, value of $K \in \{4, 8, 16, 32\}$, and class-count profile. We report $N \in \{32, 64, 96\}$. Balanced episodes assign N/K images to each selected class. Skewed episodes first assign one image to every class and distribute the remaining positions with geometrically decreasing class weights, subject to the available distinct images. Every image is evaluated.

For every reported metric, we first average episode values within each dataset and then give equal weight to the dataset means. Changes relative to frozen CLIP use the same episodes, and context contrasts use the same target queries across conditions. Pointwise 95% confidence intervals are the 2.5th and 97.5th percentiles of 10,000 bootstrap replicates of these dataset-weighted means. Each replicate resamples episodes with replacement within each dataset; for context interventions, it resamples target classes while keeping the conditions for a target together. The intervals describe variation under this sampling design for the evaluated datasets and checkpoints.

Locked episode manifests, feature-cache identifiers, method settings, task-level predictions, and source hashes are retained with the experiment records. We verify that frozen query predictions are identical across matched contexts, that prediction indices agree with their manifests, and that the decomposition in Equation 5 holds numerically before rendering the tables.

C ADDITIONAL NATIVE EPISODIC RESULTS

The following tables report every native regime for OpenAI ViT-B/32, LAION-2B ViT-B/32, and OpenAI ViT-B/16. Each entry is a five-dataset mean +[pointwise 95% confidence interval]. Full, within-support, support-only, precision, and recall values are percentages; Δ_{rank} and Δ_R are percentage points. All metrics use the complete realized episode.

Table 5: OpenAI ViT-B/32, MOON-native $N = 64$, $K = 1$–4; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.72 [61.21, 62.24]	96.01 [95.78, 96.25]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.72 [61.21, 62.24]	26.30 [25.89, 26.70]	90.69 [90.14, 91.22]
Pseudo-linear	62.25 [61.72, 62.78]	94.76 [94.47, 95.04]	−1.26 [−1.37, −1.14]	+1.79 [+1.65, +1.92]	61.97 [61.44, 62.49]	26.90 [26.48, 27.31]	89.20 [88.63, 89.75]
ZLaP	22.45 [22.15, 22.74]	94.83 [94.53, 95.12]	−1.18 [−1.36, −1.01]	−38.10 [−38.61, −37.59]	60.22 [59.61, 60.82]	7.61 [7.51, 7.72]	81.87 [81.19, 82.58]
GDA-CLIP	31.36 [31.06, 31.66]	91.41 [91.13, 91.69]	−4.60 [−4.77, −4.44]	−25.76 [−26.15, −25.38]	62.64 [62.10, 63.17]	6.89 [6.81, 6.97]	89.77 [89.19, 90.32]
TransCLIP	36.20 [35.87, 36.53]	90.46 [90.03, 90.89]	−5.52 [−5.86, −5.17]	−20.35 [−20.86, −19.84]	62.07 [61.56, 62.58]	7.46 [7.38, 7.54]	96.03 [95.65, 96.41]
ADAPT	56.84 [56.34, 57.33]	91.60 [91.20, 92.01]	−4.41 [−4.74, −4.09]	−0.47 [−0.85, −0.09]	61.72 [61.21, 62.24]	26.30 [25.89, 26.71]	90.69 [90.13, 91.21]
StatA	69.32 [68.71, 69.92]	94.90 [94.61, 95.18]	−1.11 [−1.24, −0.99]	+8.71 [+8.40, +9.02]	67.42 [66.83, 68.00]	42.51 [41.92, 43.11]	82.02 [81.36, 82.69]
MOON	72.04 [71.43, 72.66]	92.92 [92.57, 93.27]	−3.09 [−3.31, −2.87]	+13.41 [+13.06, +13.76]	70.89 [70.28, 71.51]	61.30 [60.60, 62.00]	79.81 [79.15, 80.47]
MOON (soft)	72.47 [71.86, 73.07]	92.96 [92.62, 93.31]	−3.05 [−3.27, −2.83]	+13.79 [+13.45, +14.13]	71.50 [70.90, 72.10]	64.97 [64.30, 65.65]	81.46 [80.79, 82.12]

Table 6: OpenAI ViT-B/32, MOON-native $N = 64$, $K = 2$–10; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.91 [61.56, 62.27]	90.20 [89.94, 90.46]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.91 [61.56, 62.27]	35.68 [35.34, 36.04]	87.92 [87.56, 88.29]
Pseudo-linear	62.22 [61.87, 62.59]	88.03 [87.73, 88.32]	−2.17 [−2.28, −2.05]	+2.48 [+2.36, +2.59]	62.09 [61.74, 62.45]	36.05 [35.71, 36.41]	86.68 [86.31, 87.06]
ZLaP	30.39 [30.12, 30.67]	88.37 [88.07, 88.66]	−1.84 [−1.99, −1.69]	−29.68 [−30.07, −29.30]	59.99 [59.57, 60.40]	15.00 [14.81, 15.19]	79.61 [79.14, 80.08]
GDA-CLIP	32.33 [32.12, 32.54]	83.21 [82.94, 83.48]	−6.99 [−7.13, −6.84]	−22.60 [−22.91, −22.30]	61.84 [61.47, 62.21]	13.27 [13.12, 13.42]	85.55 [85.15, 85.95]
TransCLIP	36.70 [36.48, 36.92]	78.17 [77.68, 78.63]	−12.04 [−12.43, −11.65]	−13.18 [−13.64, −12.70]	61.93 [61.57, 62.27]	14.38 [14.22, 14.54]	94.07 [93.80, 94.34]
ADAPT	60.23 [59.89, 60.57]	84.08 [83.73, 84.44]	−6.12 [−6.39, −5.84]	+4.43 [+4.13, +4.73]	61.91 [61.56, 62.27]	35.69 [35.35, 36.05]	87.91 [87.55, 88.27]
StatA	68.09 [67.70, 68.49]	87.73 [87.41, 88.04]	−2.47 [−2.61, −2.33]	+8.64 [+8.42, +8.86]	67.01 [66.63, 67.40]	48.79 [48.38, 49.22]	80.20 [79.78, 80.64]
MOON	70.95 [70.53, 71.39]	84.60 [84.24, 84.95]	−5.60 [−5.81, −5.39]	+14.64 [+14.36, +14.92]	70.22 [69.80, 70.64]	66.14 [65.66, 66.62]	76.75 [76.32, 77.19]
MOON (soft)	71.18 [70.75, 71.61]	84.61 [84.25, 84.96]	−5.59 [−5.80, −5.38]	+14.86 [+14.58, +15.13]	70.66 [70.24, 71.08]	70.09 [69.62, 70.55]	76.23 [75.79, 76.67]

Table 7: OpenAI ViT-B/32, MOON-native $N = 64$, $K = 5$–25; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.89 [61.63, 62.13]	83.00 [82.75, 83.25]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.89 [61.63, 62.13]	50.18 [49.85, 50.49]	83.43 [83.15, 83.70]
Pseudo-linear	61.99 [61.73, 62.24]	80.47 [80.21, 80.73]	−2.53 [−2.63, −2.43]	+2.63 [+2.53, +2.73]	61.94 [61.68, 62.18]	50.27 [49.95, 50.59]	82.57 [82.30, 82.85]
ZLaP	40.02 [39.76, 40.27]	81.07 [80.82, 81.33]	−1.93 [−2.04, −1.82]	−19.94 [−20.22, −19.65]	59.15 [58.87, 59.44]	30.29 [29.97, 30.60]	75.72 [75.41, 76.02]
GDA-CLIP	37.21 [37.02, 37.40]	76.80 [76.56, 77.02]	−6.20 [−6.34, −6.07]	−18.47 [−18.72, −18.22]	59.91 [59.63, 60.18]	26.45 [26.19, 26.72]	79.89 [79.60, 80.19]
TransCLIP	40.13 [39.95, 40.31]	67.08 [66.74, 67.43]	−15.94 [−16.23, −15.65]	−5.84 [−6.19, −5.49]	61.90 [61.65, 62.16]	28.52 [28.25, 28.79]	90.29 [90.07, 90.51]
ADAPT	61.68 [61.43, 61.93]	76.98 [76.71, 77.26]	−6.02 [−6.20, −5.83]	+5.81 [+5.62, +6.00]	61.89 [61.63, 62.13]	50.18 [49.86, 50.49]	83.41 [83.14, 83.69]
StatA	64.84 [64.55, 65.12]	79.35 [79.07, 79.62]	−3.66 [−3.79, −3.52]	+6.60 [+6.43, +6.78]	64.81 [64.52, 65.09]	58.36 [58.03, 58.67]	76.34 [76.05, 76.63]
MOON	66.87 [66.57, 67.18]	75.83 [75.54, 76.12]	−7.17 [−7.34, −7.00]	+12.16 [+11.96, +12.36]	66.73 [66.42, 67.03]	70.94 [70.60, 71.28]	71.42 [71.13, 71.72]
MOON (soft)	66.21 [65.89, 66.52]	75.05 [74.75, 75.35]	−7.95 [−8.13, −7.77]	+12.27 [+12.05, +12.49]	66.39 [66.08, 66.70]	75.25 [74.91, 75.58]	68.54 [68.23, 68.85]

Table 8: OpenAI ViT-B/32, MOON-native $N = 1{,}000$, $K = 5$–25; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	62.14 [61.93, 62.35]	82.67 [82.45, 82.90]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.14 [61.93, 62.35]	28.00 [27.78, 28.23]	95.34 [95.18, 95.50]
Pseudo-linear	65.37 [65.13, 65.60]	81.22 [80.99, 81.44]	−1.46 [−1.51, −1.41]	+4.69 [+4.62, +4.76]	63.14 [62.91, 63.35]	32.62 [32.39, 32.86]	88.57 [88.35, 88.79]
ZLaP	38.76 [38.53, 39.00]	82.20 [81.98, 82.42]	−0.47 [−0.53, −0.41]	−22.90 [−23.15, −22.65]	62.35 [62.14, 62.57]	20.08 [19.83, 20.32]	93.92 [93.72, 94.12]
GDA-CLIP	46.68 [46.50, 46.87]	80.88 [80.68, 81.07]	−1.80 [−1.89, −1.70]	−13.66 [−13.82, −13.50]	62.21 [62.00, 62.42]	19.41 [19.18, 19.65]	98.03 [97.90, 98.15]
TransCLIP	55.71 [55.52, 55.90]	82.42 [82.20, 82.63]	−0.25 [−0.34, −0.17]	−6.18 [−6.31, −6.04]	62.14 [61.93, 62.35]	19.33 [19.10, 19.57]	99.69 [99.64, 99.74]
ADAPT	56.35 [56.16, 56.54]	82.60 [82.38, 82.81]	−0.08 [−0.19, +0.04]	−5.71 [−5.85, −5.57]	62.14 [61.93, 62.35]	28.01 [27.79, 28.24]	95.27 [95.10, 95.43]
StatA	67.41 [67.17, 67.65]	80.22 [79.98, 80.46]	−2.45 [−2.55, −2.35]	+7.72 [+7.61, +7.84]	65.12 [64.89, 65.35]	40.38 [40.12, 40.65]	84.95 [84.71, 85.19]
MOON	68.99 [68.74, 69.25]	78.12 [77.88, 78.36]	−4.55 [−4.67, −4.43]	+11.40 [+11.27, +11.54]	67.45 [67.20, 67.70]	59.15 [58.83, 59.47]	81.45 [81.20, 81.70]
MOON (soft)	68.90 [68.65, 69.14]	78.69 [78.46, 78.92]	−3.98 [−4.09, −3.87]	+10.74 [+10.61, +10.86]	67.41 [67.17, 67.66]	61.72 [61.39, 62.04]	82.97 [82.73, 83.20]

Table 9: OpenAI ViT-B/32, MOON-native $N = 1{,}000$, $K = 25$–50; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.98 [61.88, 62.10]	73.19 [73.07, 73.31]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.98 [61.88, 62.10]	51.04 [50.84, 51.24]	96.47 [96.40, 96.55]
Pseudo-linear	63.64 [63.52, 63.76]	72.30 [72.18, 72.43]	−0.88 [−0.91, −0.86]	+2.54 [+2.50, +2.57]	62.43 [62.32, 62.54]	54.16 [53.96, 54.36]	88.28 [88.16, 88.39]
ZLaP	51.74 [51.61, 51.86]	73.08 [72.96, 73.20]	−0.11 [−0.15, −0.07]	−10.14 [−10.25, −10.04]	62.30 [62.19, 62.41]	45.34 [45.12, 45.55]	96.13 [96.05, 96.21]
GDA-CLIP	56.01 [55.91, 56.11]	74.24 [74.12, 74.36]	+1.06 [+1.01, +1.10]	−7.03 [−7.10, −6.95]	62.00 [61.89, 62.11]	44.67 [44.46, 44.89]	99.26 [99.22, 99.30]
TransCLIP	60.37 [60.27, 60.48]	74.36 [74.23, 74.48]	+1.17 [+1.12, +1.22]	−2.78 [−2.83, −2.73]	61.98 [61.88, 62.10]	44.49 [44.28, 44.70]	99.83 [99.81, 99.85]
ADAPT	60.46 [60.36, 60.56]	74.85 [74.73, 74.97]	+1.66 [+1.62, +1.70]	−3.19 [−3.24, −3.14]	61.98 [61.88, 62.10]	51.05 [50.85, 51.25]	96.45 [96.37, 96.52]
StatA	65.91 [65.79, 66.04]	71.84 [71.71, 71.97]	−1.35 [−1.41, −1.29]	+5.28 [+5.22, +5.33]	63.71 [63.59, 63.83]	61.63 [61.42, 61.83]	85.49 [85.36, 85.62]
MOON	66.14 [66.02, 66.27]	70.65 [70.53, 70.78]	−2.53 [−2.60, −2.47]	+6.69 [+6.63, +6.75]	64.86 [64.74, 64.99]	73.87 [73.68, 74.06]	81.28 [81.15, 81.41]
MOON (soft)	66.04 [65.92, 66.17]	71.13 [71.01, 71.25]	−2.05 [−2.11, −2.00]	+6.11 [+6.06, +6.17]	64.89 [64.77, 65.01]	76.44 [76.25, 76.62]	82.60 [82.47, 82.73]

Table 10: OpenAI ViT-B/32, MOON-native $N = 1{,}000$, $K = 50$–100; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.94 [61.87, 62.00]	66.77 [66.68, 66.86]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.94 [61.87, 62.00]	76.58 [76.29, 76.86]	96.30 [96.26, 96.35]
Pseudo-linear	62.36 [62.29, 62.43]	66.27 [66.18, 66.36]	−0.50 [−0.52, −0.48]	+0.92 [+0.90, +0.95]	62.03 [61.96, 62.10]	78.15 [77.87, 78.42]	87.99 [87.93, 88.06]
ZLaP	58.03 [57.95, 58.11]	66.61 [66.52, 66.70]	−0.16 [−0.20, −0.13]	−3.74 [−3.81, −3.68]	62.22 [62.16, 62.29]	73.15 [72.83, 73.47]	95.97 [95.92, 96.02]
GDA-CLIP	60.87 [60.80, 60.94]	68.57 [68.49, 68.66]	+1.80 [+1.77, +1.84]	−2.87 [−2.92, −2.82]	61.95 [61.88, 62.01]	72.53 [72.21, 72.85]	99.46 [99.44, 99.49]
TransCLIP	62.45 [62.39, 62.52]	68.30 [68.21, 68.39]	+1.53 [+1.49, +1.56]	−1.01 [−1.05, −0.98]	61.94 [61.87, 62.00]	72.40 [72.08, 72.72]	99.81 [99.79, 99.82]
ADAPT	62.85 [62.78, 62.92]	68.61 [68.52, 68.70]	+1.84 [+1.81, +1.86]	−0.93 [−0.95, −0.90]	61.94 [61.87, 62.00]	76.58 [76.29, 76.86]	96.29 [96.24, 96.33]
StatA	63.91 [63.83, 63.99]	66.10 [66.01, 66.19]	−0.67 [−0.71, −0.63]	+2.64 [+2.60, +2.69]	62.48 [62.40, 62.55]	83.30 [83.06, 83.54]	83.66 [83.58, 83.74]
MOON	63.60 [63.52, 63.68]	65.32 [65.23, 65.40]	−1.45 [−1.49, −1.42]	+3.12 [+3.07, +3.16]	62.77 [62.69, 62.84]	88.67 [88.48, 88.86]	79.19 [79.11, 79.27]
MOON (soft)	63.11 [63.02, 63.19]	64.98 [64.89, 65.07]	−1.79 [−1.82, −1.76]	+2.96 [+2.92, +3.00]	62.47 [62.39, 62.54]	90.06 [89.88, 90.24]	79.37 [79.28, 79.45]

Table 11: LAION ViT-B/32, MOON-native $N = 64$, $K = 1$–4; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.43 [65.93, 66.91]	96.67 [96.46, 96.87]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.43 [65.93, 66.91]	28.51 [28.09, 28.92]	93.00 [92.53, 93.45]
Pseudo-linear	66.54 [66.03, 67.02]	96.17 [95.93, 96.40]	−0.50 [−0.57, −0.43]	+0.61 [+0.53, +0.68]	66.48 [65.98, 66.96]	28.63 [28.20, 29.05]	92.32 [91.84, 92.79]
ZLaP	29.41 [29.08, 29.73]	95.86 [95.59, 96.13]	−0.81 [−0.96, −0.65]	−36.22 [−36.67, −35.76]	66.05 [65.49, 66.58]	8.58 [8.47, 8.69]	88.26 [87.66, 88.85]
GDA-CLIP	37.48 [37.16, 37.79]	90.00 [89.71, 90.30]	−6.67 [−6.86, −6.47]	−22.28 [−22.67, −21.89]	67.56 [67.05, 68.05]	7.69 [7.61, 7.77]	93.00 [92.52, 93.47]
TransCLIP	42.64 [42.27, 42.99]	95.16 [94.87, 95.44]	−1.51 [−1.70, −1.31]	−22.32 [−22.68, −21.96]	66.47 [65.97, 66.96]	8.63 [8.54, 8.73]	97.87 [97.61, 98.14]
ADAPT	62.65 [62.18, 63.12]	93.34 [92.98, 93.70]	−3.33 [−3.61, −3.05]	−0.45 [−0.78, −0.12]	66.43 [65.93, 66.91]	28.51 [28.09, 28.92]	93.00 [92.53, 93.45]
StatA	71.27 [70.71, 71.82]	96.39 [96.15, 96.61]	−0.28 [−0.35, −0.21]	+5.12 [+4.91, +5.33]	69.50 [68.96, 70.01]	37.22 [36.66, 37.77]	86.72 [86.14, 87.31]
MOON	76.15 [75.57, 76.71]	95.21 [94.94, 95.49]	−1.46 [−1.60, −1.32]	+11.17 [+10.90, +11.45]	74.23 [73.66, 74.79]	56.21 [55.54, 56.87]	84.59 [84.00, 85.20]
MOON (soft)	76.85 [76.28, 77.41]	95.20 [94.92, 95.48]	−1.47 [−1.60, −1.33]	+11.89 [+11.63, +12.16]	75.18 [74.62, 75.73]	60.92 [60.26, 61.58]	85.67 [85.07, 86.27]

Table 12: LAION ViT-B/32, MOON-native $N = 64$, $K = 2$–10; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.69 [66.37, 67.03]	91.96 [91.72, 92.19]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.69 [66.37, 67.03]	38.55 [38.19, 38.92]	90.60 [90.27, 90.94]
Pseudo-linear	66.79 [66.46, 67.12]	90.96 [90.70, 91.21]	−1.00 [−1.08, −0.91]	+1.09 [+1.01, +1.17]	66.74 [66.41, 67.07]	38.61 [38.25, 38.98]	90.11 [89.76, 90.45]
ZLaP	37.91 [37.60, 38.22]	90.81 [90.53, 91.07]	−1.15 [−1.29, −1.01]	−27.63 [−27.99, −27.27]	65.96 [65.59, 66.32]	17.24 [17.01, 17.46]	85.61 [85.22, 85.99]
GDA-CLIP	37.40 [37.18, 37.62]	82.14 [81.88, 82.40]	−9.81 [−9.97, −9.65]	−19.48 [−19.80, −19.17]	66.61 [66.27, 66.95]	14.34 [14.18, 14.49]	88.74 [88.37, 89.09]
TransCLIP	41.96 [41.71, 42.20]	86.35 [85.97, 86.73]	−5.60 [−5.91, −5.30]	−19.13 [−19.51, −18.77]	66.69 [66.37, 67.03]	15.91 [15.74, 16.07]	95.91 [95.69, 96.13]
ADAPT	65.30 [64.97, 65.62]	87.06 [86.73, 87.40]	−4.89 [−5.14, −4.65]	+3.50 [+3.22, +3.77]	66.69 [66.37, 67.03]	38.55 [38.20, 38.92]	90.60 [90.27, 90.94]
StatA	70.77 [70.40, 71.15]	91.34 [91.08, 91.60]	−0.61 [−0.70, −0.52]	+4.69 [+4.53, +4.85]	69.61 [69.25, 69.97]	45.42 [45.02, 45.82]	85.73 [85.34, 86.12]
MOON	75.57 [75.19, 75.96]	89.29 [89.00, 89.57]	−2.67 [−2.82, −2.53]	+11.55 [+11.32, +11.77]	74.24 [73.87, 74.62]	63.35 [62.89, 63.81]	83.08 [82.69, 83.47]
MOON (soft)	76.40 [76.02, 76.79]	89.09 [88.80, 89.38]	−2.86 [−3.01, −2.72]	+12.57 [+12.34, +12.80]	75.32 [74.95, 75.70]	68.83 [68.37, 69.29]	82.29 [81.90, 82.69]

Table 13: LAION ViT-B/32, MOON-native $N = 64$, $K = 5$–25; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.83 [66.60, 67.06]	85.81 [85.58, 86.03]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.83 [66.60, 67.06]	53.71 [53.39, 54.03]	86.25 [86.00, 86.49]
Pseudo-linear	66.86 [66.63, 67.09]	84.66 [84.43, 84.89]	−1.15 [−1.22, −1.07]	+1.18 [+1.10, +1.25]	66.85 [66.62, 67.08]	53.72 [53.40, 54.04]	86.06 [85.81, 86.30]
ZLaP	47.58 [47.32, 47.84]	84.53 [84.29, 84.76]	−1.29 [−1.39, −1.18]	−17.97 [−18.22, −17.71]	65.26 [64.99, 65.52]	34.25 [33.91, 34.59]	81.31 [81.04, 81.59]
GDA-CLIP	40.94 [40.75, 41.13]	77.68 [77.47, 77.89]	−8.13 [−8.27, −8.00]	−17.76 [−18.00, −17.52]	64.40 [64.14, 64.65]	27.66 [27.39, 27.93]	82.55 [82.26, 82.83]
TransCLIP	44.53 [44.35, 44.72]	74.96 [74.58, 75.33]	−10.85 [−11.17, −10.54]	−11.45 [−11.81, −11.09]	66.81 [66.58, 67.05]	30.48 [30.21, 30.75]	92.35 [92.15, 92.55]
ADAPT	66.62 [66.39, 66.85]	80.81 [80.56, 81.06]	−5.00 [−5.17, −4.84]	+4.79 [+4.62, +4.96]	66.83 [66.60, 67.06]	53.71 [53.39, 54.03]	86.25 [86.00, 86.49]
StatA	69.12 [68.87, 69.37]	84.79 [84.54, 85.03]	−1.02 [−1.11, −0.93]	+3.31 [+3.19, +3.43]	68.76 [68.51, 69.00]	57.77 [57.45, 58.09]	82.28 [82.02, 82.53]
MOON	72.06 [71.78, 72.33]	82.12 [81.86, 82.37]	−3.70 [−3.82, −3.57]	+8.92 [+8.75, +9.09]	71.51 [71.23, 71.77]	69.95 [69.63, 70.27]	78.28 [78.01, 78.54]
MOON (soft)	72.28 [72.00, 72.56]	81.51 [81.25, 81.77]	−4.30 [−4.43, −4.17]	+9.75 [+9.57, +9.93]	72.01 [71.73, 72.28]	74.93 [74.62, 75.26]	75.88 [75.59, 76.15]

Table 14: LAION ViT-B/32, MOON-native $N = 1{,}000$, $K = 5$–25; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.86 [66.67, 67.04]	85.43 [85.24, 85.63]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.86 [66.67, 67.04]	29.94 [29.71, 30.17]	96.60 [96.46, 96.74]
Pseudo-linear	69.08 [68.88, 69.28]	84.88 [84.68, 85.08]	−0.55 [−0.59, −0.51]	+2.78 [+2.72, +2.83]	67.36 [67.16, 67.55]	32.03 [31.80, 32.27]	91.34 [91.15, 91.53]
ZLaP	45.29 [45.07, 45.52]	85.48 [85.28, 85.68]	+0.05 [−0.01, +0.10]	−21.61 [−21.83, −21.39]	66.87 [66.67, 67.06]	20.28 [20.04, 20.52]	95.81 [95.64, 95.99]
GDA-CLIP	53.31 [53.13, 53.48]	81.78 [81.61, 81.95]	−3.65 [−3.75, −3.55]	−9.90 [−10.03, −9.77]	66.95 [66.76, 67.14]	19.70 [19.46, 19.94]	99.16 [99.08, 99.25]
TransCLIP	60.72 [60.53, 60.90]	86.09 [85.90, 86.28]	+0.66 [+0.58, +0.74]	−6.80 [−6.92, −6.68]	66.86 [66.67, 67.04]	19.79 [19.56, 20.03]	99.76 [99.72, 99.80]
ADAPT	62.33 [62.15, 62.52]	85.89 [85.69, 86.08]	+0.45 [+0.35, +0.55]	−4.98 [−5.09, −4.86]	66.86 [66.67, 67.04]	29.94 [29.72, 30.17]	96.59 [96.45, 96.72]
StatA	70.08 [69.87, 70.28]	85.11 [84.90, 85.31]	−0.32 [−0.38, −0.26]	+3.55 [+3.47, +3.63]	68.21 [68.01, 68.41]	34.99 [34.74, 35.24]	89.65 [89.44, 89.86]
MOON	73.60 [73.38, 73.81]	83.03 [82.82, 83.23]	−2.41 [−2.50, −2.32]	+9.15 [+9.04, +9.26]	71.23 [71.01, 71.44]	54.23 [53.93, 54.52]	86.14 [85.91, 86.36]
MOON (soft)	73.77 [73.55, 73.98]	83.23 [83.02, 83.43]	−2.21 [−2.29, −2.13]	+9.12 [+9.01, +9.22]	71.57 [71.36, 71.78]	58.13 [57.82, 58.43]	86.92 [86.69, 87.13]

Table 15: LAION ViT-B/32, MOON-native $N = 1{,}000$, $K = 25$–50; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.88 [66.78, 66.98]	77.23 [77.12, 77.35]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.88 [66.78, 66.98]	52.51 [52.31, 52.71]	97.40 [97.33, 97.46]
Pseudo-linear	68.23 [68.12, 68.34]	76.98 [76.87, 77.09]	−0.25 [−0.27, −0.23]	+1.60 [+1.57, +1.63]	67.08 [66.98, 67.18]	53.97 [53.77, 54.17]	91.64 [91.54, 91.75]
ZLaP	57.37 [57.25, 57.49]	77.60 [77.49, 77.71]	+0.36 [+0.33, +0.40]	−9.88 [−9.97, −9.78]	66.90 [66.80, 67.01]	45.55 [45.34, 45.77]	97.10 [97.03, 97.17]
GDA-CLIP	61.17 [61.08, 61.27]	77.16 [77.05, 77.27]	−0.08 [−0.12, −0.03]	−5.63 [−5.69, −5.57]	66.89 [66.79, 66.99]	44.72 [44.51, 44.94]	99.68 [99.66, 99.71]
TransCLIP	65.29 [65.20, 65.39]	78.81 [78.70, 78.93]	+1.58 [+1.54, +1.62]	−3.17 [−3.21, −3.12]	66.88 [66.78, 66.98]	44.70 [44.49, 44.91]	99.86 [99.84, 99.87]
ADAPT	65.58 [65.48, 65.68]	78.88 [78.77, 79.00]	+1.65 [+1.61, +1.68]	−2.95 [−3.00, −2.91]	66.88 [66.78, 66.98]	52.52 [52.32, 52.72]	97.38 [97.32, 97.44]
StatA	69.99 [69.88, 70.10]	77.61 [77.50, 77.73]	+0.38 [+0.34, +0.42]	+2.73 [+2.69, +2.77]	67.72 [67.62, 67.83]	57.19 [56.99, 57.39]	90.74 [90.63, 90.85]
MOON	71.48 [71.37, 71.59]	76.26 [76.15, 76.37]	−0.97 [−1.02, −0.93]	+5.57 [+5.52, +5.62]	69.48 [69.37, 69.59]	71.69 [71.50, 71.89]	86.40 [86.29, 86.52]
MOON (soft)	71.54 [71.43, 71.65]	76.44 [76.33, 76.55]	−0.79 [−0.84, −0.75]	+5.45 [+5.41, +5.50]	69.69 [69.59, 69.80]	74.89 [74.70, 75.08]	87.17 [87.06, 87.29]

Table 16: LAION ViT-B/32, MOON-native $N = 1{,}000$, $K = 50$–100; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.88 [66.82, 66.94]	71.38 [71.29, 71.46]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.88 [66.82, 66.94]	77.70 [77.42, 77.97]	97.20 [97.16, 97.24]
Pseudo-linear	67.38 [67.32, 67.44]	71.32 [71.24, 71.41]	−0.05 [−0.07, −0.04]	+0.55 [+0.54, +0.57]	66.90 [66.84, 66.96]	78.46 [78.19, 78.73]	91.79 [91.73, 91.85]
ZLaP	63.58 [63.49, 63.65]	71.76 [71.68, 71.85]	+0.39 [+0.36, +0.42]	−3.70 [−3.76, −3.63]	66.87 [66.80, 66.93]	73.40 [73.09, 73.72]	96.82 [96.78, 96.87]
GDA-CLIP	65.51 [65.44, 65.58]	72.41 [72.34, 72.49]	+1.04 [+1.01, +1.07]	−2.41 [−2.45, −2.37]	66.89 [66.83, 66.94]	72.54 [72.23, 72.87]	99.81 [99.79, 99.82]
TransCLIP	67.21 [67.15, 67.28]	73.00 [72.91, 73.08]	+1.62 [+1.59, +1.65]	−1.29 [−1.32, −1.26]	66.88 [66.82, 66.94]	72.49 [72.18, 72.82]	99.85 [99.84, 99.86]
ADAPT	67.50 [67.44, 67.56]	72.87 [72.78, 72.96]	+1.50 [+1.47, +1.52]	−0.88 [−0.90, −0.85]	66.88 [66.82, 66.94]	77.70 [77.42, 77.98]	97.18 [97.14, 97.22]
StatA	69.26 [69.19, 69.32]	72.26 [72.18, 72.35]	+0.89 [+0.86, +0.92]	+1.49 [+1.46, +1.52]	67.33 [67.27, 67.40]	80.85 [80.59, 81.10]	90.21 [90.15, 90.27]
MOON	69.34 [69.27, 69.41]	71.21 [71.12, 71.29]	−0.17 [−0.20, −0.14]	+2.63 [+2.59, +2.67]	67.96 [67.89, 68.03]	88.38 [88.19, 88.57]	84.67 [84.60, 84.74]
MOON (soft)	69.00 [68.92, 69.07]	70.77 [70.68, 70.85]	−0.61 [−0.64, −0.58]	+2.72 [+2.69, +2.76]	67.81 [67.74, 67.88]	90.29 [90.11, 90.47]	84.22 [84.14, 84.31]

Table 17: OpenAI ViT-B/16, MOON-native $N = 64$, $K = 1$–4; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.29 [64.77, 65.81]	96.87 [96.66, 97.07]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.29 [64.77, 65.81]	31.68 [31.19, 32.17]	91.05 [90.52, 91.57]
Pseudo-linear	66.13 [65.60, 66.66]	95.63 [95.37, 95.89]	−1.23 [−1.35, −1.12]	+2.06 [+1.92, +2.21]	65.72 [65.19, 66.25]	32.92 [32.40, 33.43]	89.81 [89.24, 90.36]
ZLaP	26.33 [26.01, 26.65]	95.82 [95.55, 96.09]	−1.04 [−1.22, −0.88]	−37.92 [−38.43, −37.40]	63.89 [63.30, 64.49]	8.62 [8.51, 8.73]	84.79 [84.14, 85.46]
GDA-CLIP	34.19 [33.86, 34.50]	93.13 [92.90, 93.36]	−3.74 [−3.88, −3.59]	−27.37 [−27.76, −26.98]	66.32 [65.78, 66.86]	7.38 [7.30, 7.46]	91.83 [91.33, 92.34]
TransCLIP	39.15 [38.81, 39.48]	92.22 [91.83, 92.61]	−4.62 [−4.95, −4.29]	−21.74 [−22.25, −21.24]	65.54 [65.01, 66.06]	8.06 [7.98, 8.15]	97.39 [97.09, 97.68]
ADAPT	60.76 [60.25, 61.26]	92.08 [91.67, 92.49]	−4.78 [−5.13, −4.44]	+0.25 [−0.15, +0.64]	65.29 [64.77, 65.81]	31.68 [31.20, 32.17]	91.03 [90.50, 91.56]
StatA	72.55 [71.91, 73.18]	95.75 [95.49, 96.01]	−1.11 [−1.23, −0.99]	+8.37 [+8.05, +8.69]	70.59 [69.99, 71.20]	49.40 [48.76, 50.03]	82.33 [81.65, 83.01]
MOON	74.49 [73.86, 75.13]	93.87 [93.54, 94.19]	−3.00 [−3.22, −2.79]	+12.20 [+11.86, +12.54]	73.34 [72.72, 73.97]	65.11 [64.42, 65.80]	81.24 [80.58, 81.92]
MOON (soft)	74.74 [74.13, 75.36]	93.90 [93.57, 94.23]	−2.97 [−3.18, −2.76]	+12.42 [+12.10, +12.74]	73.58 [72.99, 74.19]	67.49 [66.82, 68.15]	82.95 [82.29, 83.62]

Table 18: OpenAI ViT-B/16, MOON-native $N = 64$, $K = 2$–10; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.65 [65.30, 66.01]	92.19 [91.94, 92.43]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.65 [65.30, 66.01]	40.34 [39.96, 40.72]	88.39 [88.02, 88.75]
Pseudo-linear	66.21 [65.84, 66.57]	89.95 [89.67, 90.23]	−2.24 [−2.36, −2.11]	+2.79 [+2.66, +2.91]	65.97 [65.61, 66.33]	41.12 [40.73, 41.51]	87.37 [86.98, 87.74]
ZLaP	34.80 [34.50, 35.09]	90.87 [90.60, 91.14]	−1.32 [−1.46, −1.18]	−29.54 [−29.92, −29.15]	64.27 [63.87, 64.66]	16.72 [16.51, 16.94]	82.14 [81.70, 82.59]
GDA-CLIP	34.93 [34.71, 35.15]	86.51 [86.27, 86.75]	−5.68 [−5.83, −5.52]	−25.04 [−25.35, −24.75]	65.56 [65.19, 65.93]	13.94 [13.78, 14.10]	86.86 [86.47, 87.24]
TransCLIP	39.53 [39.29, 39.76]	81.63 [81.17, 82.08]	−10.56 [−10.96, −10.18]	−15.57 [−16.04, −15.09]	65.66 [65.30, 66.01]	15.16 [15.00, 15.32]	94.91 [94.66, 95.15]
ADAPT	63.88 [63.53, 64.22]	85.76 [85.39, 86.11]	−6.43 [−6.73, −6.14]	+4.66 [+4.34, +4.98]	65.65 [65.30, 66.01]	40.34 [39.96, 40.73]	88.38 [88.01, 88.74]
StatA	71.36 [70.94, 71.78]	89.59 [89.28, 89.88]	−2.60 [−2.74, −2.46]	+8.31 [+8.07, +8.54]	70.34 [69.94, 70.75]	55.19 [54.74, 55.64]	80.24 [79.81, 80.69]
MOON	73.72 [73.30, 74.16]	86.64 [86.30, 86.97]	−5.55 [−5.76, −5.34]	+13.62 [+13.34, +13.89]	73.10 [72.69, 73.53]	69.97 [69.50, 70.46]	78.23 [77.80, 78.65]
MOON (soft)	73.78 [73.35, 74.21]	86.50 [86.14, 86.83]	−5.69 [−5.90, −5.48]	+13.81 [+13.54, +14.08]	73.23 [72.80, 73.66]	72.91 [72.44, 73.37]	77.71 [77.26, 78.15]

Table 19: OpenAI ViT-B/16, MOON-native $N = 64$, $K = 5$–25; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.74 [65.50, 65.98]	85.79 [85.56, 86.03]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.74 [65.50, 65.98]	54.22 [53.89, 54.54]	84.24 [83.97, 84.51]
Pseudo-linear	65.98 [65.73, 66.22]	83.35 [83.10, 83.60]	−2.44 [−2.54, −2.34]	+2.67 [+2.57, +2.78]	65.91 [65.67, 66.15]	54.56 [54.24, 54.88]	83.67 [83.40, 83.94]
ZLaP	44.59 [44.33, 44.84]	84.26 [84.02, 84.50]	−1.53 [−1.64, −1.42]	−19.63 [−19.90, −19.35]	63.55 [63.27, 63.83]	32.94 [32.61, 33.27]	78.31 [78.01, 78.61]
GDA-CLIP	39.72 [39.53, 39.91]	80.85 [80.63, 81.07]	−4.94 [−5.08, −4.80]	−21.08 [−21.32, −20.83]	63.68 [63.42, 63.94]	27.43 [27.15, 27.70]	81.06 [80.77, 81.34]
TransCLIP	42.76 [42.58, 42.93]	70.70 [70.36, 71.04]	−15.10 [−15.40, −14.80]	−7.90 [−8.26, −7.54]	65.75 [65.50, 65.99]	29.44 [29.16, 29.71]	90.93 [90.71, 91.15]
ADAPT	65.44 [65.20, 65.68]	79.79 [79.53, 80.06]	−6.00 [−6.18, −5.81]	+5.70 [+5.51, +5.88]	65.74 [65.50, 65.98]	54.22 [53.89, 54.54]	84.23 [83.97, 84.50]
StatA	68.51 [68.23, 68.79]	81.90 [81.63, 82.17]	−3.89 [−4.02, −3.76]	+6.65 [+6.48, +6.82]	68.36 [68.08, 68.64]	63.41 [63.08, 63.72]	76.71 [76.42, 77.00]
MOON	69.94 [69.65, 70.23]	78.74 [78.46, 79.02]	−7.05 [−7.22, −6.89]	+11.25 [+11.05, +11.45]	69.75 [69.46, 70.04]	73.99 [73.66, 74.31]	72.82 [72.53, 73.12]
MOON (soft)	69.12 [68.82, 69.41]	77.92 [77.62, 78.22]	−7.87 [−8.05, −7.69]	+11.25 [+11.03, +11.47]	69.19 [68.89, 69.48]	77.33 [76.99, 77.66]	70.09 [69.78, 70.40]

Table 20: OpenAI ViT-B/16, MOON-native $N = 1{,}000$, $K = 5$–25; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.91 [65.70, 66.11]	85.39 [85.18, 85.61]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.91 [65.70, 66.11]	31.35 [31.12, 31.58]	94.19 [94.01, 94.36]
Pseudo-linear	69.14 [68.91, 69.37]	83.90 [83.68, 84.12]	−1.50 [−1.55, −1.44]	+4.73 [+4.66, +4.81]	66.95 [66.73, 67.16]	37.25 [37.00, 37.50]	87.90 [87.68, 88.13]
ZLaP	42.77 [42.53, 43.01]	85.41 [85.20, 85.62]	+0.01 [−0.04, +0.07]	−23.15 [−23.40, −22.90]	66.09 [65.87, 66.31]	20.75 [20.51, 20.99]	94.22 [94.02, 94.42]
GDA-CLIP	50.46 [50.27, 50.65]	84.34 [84.16, 84.51]	−1.05 [−1.16, −0.95]	−14.39 [−14.55, −14.23]	66.02 [65.81, 66.22]	19.67 [19.43, 19.91]	97.98 [97.86, 98.09]
TransCLIP	59.27 [59.08, 59.46]	85.80 [85.60, 86.00]	+0.41 [+0.30, +0.51]	−7.05 [−7.19, −6.91]	65.91 [65.70, 66.12]	19.62 [19.39, 19.86]	99.46 [99.40, 99.53]
ADAPT	60.84 [60.65, 61.04]	85.49 [85.29, 85.70]	+0.10 [−0.03, +0.22]	−5.16 [−5.31, −5.02]	65.91 [65.70, 66.12]	31.38 [31.15, 31.60]	94.17 [94.00, 94.35]
StatA	71.58 [71.34, 71.81]	82.94 [82.70, 83.17]	−2.46 [−2.56, −2.36]	+8.13 [+8.01, +8.24]	68.88 [68.65, 69.11]	45.54 [45.25, 45.81]	84.50 [84.26, 84.74]
MOON	72.41 [72.15, 72.66]	80.47 [80.24, 80.70]	−4.92 [−5.05, −4.79]	+11.42 [+11.29, +11.56]	70.74 [70.49, 70.99]	63.03 [62.70, 63.36]	82.23 [81.98, 82.47]
MOON (soft)	72.00 [71.75, 72.24]	80.93 [80.70, 81.15]	−4.47 [−4.59, −4.35]	+10.56 [+10.43, +10.69]	70.36 [70.11, 70.60]	64.13 [63.81, 64.45]	83.44 [83.21, 83.68]

Table 21: OpenAI ViT-B/16, MOON-native $N = 1{,}000$, $K = 25$–50; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.88 [65.77, 65.99]	76.65 [76.53, 76.77]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.88 [65.77, 65.99]	53.92 [53.72, 54.12]	94.99 [94.91, 95.07]
Pseudo-linear	67.65 [67.54, 67.77]	75.76 [75.64, 75.88]	−0.89 [−0.92, −0.86]	+2.67 [+2.63, +2.70]	66.39 [66.28, 66.51]	57.76 [57.56, 57.96]	87.79 [87.68, 87.91]
ZLaP	55.60 [55.48, 55.73]	76.70 [76.58, 76.83]	+0.05 [+0.01, +0.09]	−10.33 [−10.44, −10.22]	66.21 [66.10, 66.32]	46.00 [45.79, 46.21]	96.03 [95.95, 96.11]
GDA-CLIP	60.21 [60.11, 60.31]	78.62 [78.51, 78.73]	+1.97 [+1.91, +2.02]	−7.64 [−7.71, −7.56]	65.93 [65.82, 66.04]	44.85 [44.63, 45.06]	98.42 [98.36, 98.47]
TransCLIP	64.46 [64.36, 64.56]	78.49 [78.38, 78.62]	+1.84 [+1.79, +1.90]	−3.27 [−3.32, −3.21]	65.88 [65.77, 65.99]	44.64 [44.42, 44.85]	99.50 [99.47, 99.53]
ADAPT	65.00 [64.90, 65.11]	78.89 [78.77, 79.01]	+2.24 [+2.19, +2.28]	−3.11 [−3.16, −3.06]	65.88 [65.78, 65.99]	53.94 [53.74, 54.15]	94.96 [94.88, 95.04]
StatA	70.43 [70.31, 70.56]	75.38 [75.26, 75.51]	−1.27 [−1.33, −1.20]	+5.82 [+5.76, +5.88]	67.56 [67.44, 67.68]	65.32 [65.13, 65.52]	84.65 [84.53, 84.78]
MOON	69.80 [69.68, 69.93]	73.97 [73.85, 74.09]	−2.68 [−2.75, −2.62]	+6.60 [+6.54, +6.67]	68.41 [68.29, 68.54]	76.20 [76.02, 76.39]	81.93 [81.80, 82.06]
MOON (soft)	69.35 [69.23, 69.48]	74.00 [73.88, 74.13]	−2.65 [−2.71, −2.59]	+6.12 [+6.06, +6.18]	68.02 [67.90, 68.14]	77.44 [77.27, 77.62]	82.59 [82.47, 82.72]

Table 22: OpenAI ViT-B/16, MOON-native $N = 1{,}000$, $K = 50$–100; five-dataset macro. All accuracy metrics use the full realized batch. Entries are mean [95% CI]. A_{full}, A_{support}, A_{SO}, P, R are percentages; Δ_{rank} and Δ_R are percentage points.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.86 [65.79, 65.92]	70.57 [70.48, 70.65]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.86 [65.79, 65.92]	78.27 [78.00, 78.54]	94.80 [94.74, 94.85]
Pseudo-linear	66.30 [66.23, 66.37]	70.06 [69.97, 70.15]	−0.50 [−0.52, −0.48]	+0.94 [+0.92, +0.97]	65.97 [65.90, 66.03]	80.05 [79.80, 80.31]	87.29 [87.22, 87.36]
ZLaP	62.01 [61.93, 62.09]	70.31 [70.22, 70.40]	−0.26 [−0.29, −0.22]	−3.59 [−3.65, −3.52]	66.08 [66.01, 66.14]	73.73 [73.42, 74.05]	95.61 [95.56, 95.67]
GDA-CLIP	65.51 [65.43, 65.58]	73.39 [73.30, 73.47]	+2.82 [+2.78, +2.85]	−3.17 [−3.22, −3.12]	65.88 [65.82, 65.95]	72.63 [72.32, 72.95]	98.95 [98.92, 98.98]
TransCLIP	66.89 [66.83, 66.96]	72.84 [72.75, 72.93]	+2.28 [+2.24, +2.31]	−1.24 [−1.27, −1.21]	65.85 [65.79, 65.92]	72.48 [72.16, 72.80]	99.50 [99.49, 99.52]
ADAPT	67.29 [67.22, 67.35]	72.94 [72.85, 73.03]	+2.37 [+2.35, +2.40]	−0.94 [−0.97, −0.91]	65.86 [65.79, 65.92]	78.27 [78.01, 78.55]	94.78 [94.73, 94.83]
StatA	68.29 [68.21, 68.37]	70.18 [70.09, 70.26]	−0.39 [−0.43, −0.35]	+2.82 [+2.78, +2.87]	66.29 [66.22, 66.36]	85.15 [84.93, 85.38]	82.88 [82.80, 82.95]
MOON	67.29 [67.21, 67.37]	68.96 [68.88, 69.05]	−1.60 [−1.64, −1.56]	+3.03 [+2.99, +3.08]	66.39 [66.31, 66.46]	89.47 [89.28, 89.65]	79.54 [79.46, 79.63]
MOON (soft)	66.63 [66.55, 66.71]	68.36 [68.27, 68.44]	−2.21 [−2.25, −2.17]	+2.98 [+2.94, +3.03]	65.86 [65.78, 65.93]	90.23 [90.05, 90.40]	79.31 [79.22, 79.40]

D ADDITIONAL CONTEXT-INTERVENTION RESULTS

The query tables report all three context conditions and all three checkpoints. Predictions are evaluated on the same eight target queries in near, random, and far contexts, while predicted-support precision and recall use the full 64-image episode. The complete-batch tables then evaluate all 64 images. Together they separate the effect on unchanged queries from behavior over the episode as a whole.

D.1 FIXED-QUERY AND COMPLETE-BATCH RESULTS

Table 23: OpenAI ViT-B/32 matched near context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.14 [58.68, 63.56]	67.69 [65.25, 70.02]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.14 [58.68, 63.56]	48.92 [47.77, 50.10]	89.10 [88.06, 90.14]
Pseudo-linear	61.22 [58.73, 63.66]	67.70 [65.08, 70.21]	+0.01 [−0.77, +0.83]	+0.06 [−0.75, +0.82]	61.19 [58.71, 63.62]	48.55 [47.39, 49.74]	86.06 [84.96, 87.14]
ZLaP	39.00 [36.42, 41.55]	66.29 [63.79, 68.64]	−1.39 [−2.87, +0.04]	−20.75 [−22.71, −18.83]	59.74 [57.01, 62.37]	21.60 [21.09, 22.13]	78.34 [77.15, 79.54]
GDA-CLIP	43.60 [41.37, 45.81]	64.76 [62.65, 66.82]	−2.93 [−4.24, −1.58]	−14.62 [−16.11, −13.15]	60.83 [58.23, 63.38]	19.41 [19.03, 19.80]	87.31 [86.15, 88.42]
TransCLIP	44.54 [42.21, 46.85]	65.89 [63.46, 68.25]	−1.80 [−3.49, −0.08]	−14.80 [−16.46, −13.21]	61.14 [58.68, 63.56]	21.08 [20.73, 21.46]	97.36 [96.83, 97.86]
ADAPT	60.24 [57.69, 62.65]	66.22 [63.63, 68.67]	−1.47 [−2.61, −0.34]	+0.57 [−0.25, +1.36]	61.08 [58.56, 63.53]	48.83 [47.66, 50.02]	86.21 [85.09, 87.31]
StatA	61.67 [59.00, 64.25]	65.65 [62.96, 68.24]	−2.04 [−3.76, −0.39]	+2.57 [+1.57, +3.60]	62.85 [60.32, 65.30]	63.75 [62.60, 64.90]	81.41 [80.37, 82.46]
MOON	64.17 [61.38, 66.88]	66.76 [63.86, 69.55]	−0.93 [−2.13, +0.30]	+3.96 [+2.97, +4.96]	64.07 [61.38, 66.66]	72.32 [70.87, 73.76]	73.41 [72.22, 74.59]
MOON (soft)	63.95 [61.17, 66.66]	65.84 [62.94, 68.60]	−1.84 [−2.99, −0.72]	+4.65 [+3.66, +5.70]	64.17 [61.43, 66.77]	77.66 [76.09, 79.20]	73.97 [72.65, 75.25]

Table 24: OpenAI ViT-B/32 matched random context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.14 [58.68, 63.56]	87.16 [85.24, 89.01]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.14 [58.68, 63.56]	40.78 [39.85, 41.70]	88.20 [87.31, 89.06]
Pseudo-linear	61.29 [58.77, 63.74]	84.99 [82.74, 87.12]	−2.17 [−3.14, −1.22]	+2.31 [+1.37, +3.29]	61.36 [58.86, 63.80]	40.88 [39.95, 41.81]	86.95 [86.04, 87.84]
ZLaP	34.76 [32.53, 37.02]	86.14 [84.17, 88.01]	−1.02 [−1.88, −0.18]	−25.36 [−27.63, −23.07]	58.43 [55.63, 61.16]	19.54 [19.13, 19.95]	79.17 [78.08, 80.25]
GDA-CLIP	31.68 [29.74, 33.63]	78.81 [76.98, 80.61]	−8.35 [−9.74, −6.94]	−21.12 [−23.31, −18.96]	59.89 [57.22, 62.53]	16.93 [16.68, 17.19]	84.49 [83.50, 85.45]
TransCLIP	35.32 [33.34, 37.28]	80.33 [78.18, 82.49]	−6.82 [−8.35, −5.33]	−19.00 [−21.20, −16.82]	61.12 [58.64, 63.54]	18.13 [17.91, 18.34]	93.79 [93.15, 94.44]
ADAPT	62.13 [59.59, 64.66]	81.21 [78.77, 83.57]	−5.95 [−7.69, −4.24]	+6.94 [+5.21, +8.69]	61.92 [59.38, 64.44]	42.18 [41.24, 43.11]	86.60 [85.65, 87.54]
StatA	66.97 [64.31, 69.56]	83.14 [80.88, 85.34]	−4.02 [−5.29, −2.75]	+9.85 [+8.16, +11.57]	66.58 [63.93, 69.16]	54.81 [53.66, 55.96]	81.64 [80.63, 82.63]
MOON	69.53 [66.57, 72.36]	80.97 [78.27, 83.54]	−6.19 [−7.91, −4.51]	+14.58 [+12.78, +16.44]	69.41 [66.54, 72.12]	66.84 [65.57, 68.04]	76.34 [75.28, 77.37]
MOON (soft)	70.26 [67.34, 73.04]	80.67 [77.98, 83.27]	−6.49 [−8.17, −4.84]	+15.61 [+13.75, +17.49]	70.10 [67.23, 72.83]	72.25 [71.12, 73.34]	75.72 [74.63, 76.79]

Table 25: OpenAI ViT-B/32 matched far context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	61.14 [58.68, 63.56]	97.24 [96.08, 98.26]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.14 [58.68, 63.56]	49.44 [48.44, 50.47]	94.60 [94.03, 95.15]
Pseudo-linear	61.40 [58.90, 63.85]	96.23 [94.84, 97.48]	−1.00 [−1.50, −0.56]	+1.26 [+0.77, +1.78]	61.37 [58.88, 63.82]	49.87 [48.88, 50.89]	94.17 [93.57, 94.77]
ZLaP	30.94 [28.88, 32.95]	94.28 [92.59, 95.82]	−2.95 [−4.04, −1.96]	−27.24 [−29.66, −24.85]	56.87 [54.10, 59.62]	22.74 [22.34, 23.15]	88.63 [87.85, 89.41]
GDA-CLIP	24.76 [23.11, 26.41]	81.86 [80.15, 83.48]	−15.38 [−16.86, −13.94]	−21.00 [−23.63, −18.38]	58.32 [55.53, 61.05]	19.96 [19.72, 20.19]	94.02 [93.44, 94.58]
TransCLIP	29.13 [27.43, 30.82]	94.85 [93.52, 96.07]	−2.39 [−3.35, −1.45]	−29.62 [−31.81, −27.40]	60.94 [58.47, 63.36]	20.37 [20.15, 20.59]	97.63 [97.26, 98.00]
ADAPT	62.65 [60.08, 65.20]	91.14 [89.11, 93.04]	−6.10 [−7.72, −4.57]	+7.61 [+6.01, +9.28]	62.41 [59.85, 64.95]	53.67 [52.56, 54.81]	93.78 [93.11, 94.42]
StatA	69.70 [67.14, 72.21]	95.57 [94.28, 96.74]	−1.67 [−2.42, −0.96]	+10.23 [+8.53, +11.93]	69.10 [66.50, 71.64]	61.55 [60.46, 62.65]	90.18 [89.38, 90.96]
MOON	71.96 [68.94, 74.89]	93.48 [91.68, 95.10]	−3.75 [−4.98, −2.68]	+14.57 [+12.73, +16.45]	71.66 [68.64, 74.53]	78.71 [77.66, 79.76]	88.77 [87.97, 89.56]
MOON (soft)	72.79 [69.85, 75.62]	93.42 [91.65, 95.03]	−3.82 [−5.02, −2.75]	+15.46 [+13.65, +17.33]	72.55 [69.67, 75.33]	83.21 [82.26, 84.16]	88.46 [87.74, 89.17]

Table 26: OpenAI ViT-B/32 matched near context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	56.63 [55.66, 57.59]	70.89 [69.98, 71.81]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	56.63 [55.66, 57.59]	48.92 [47.77, 50.10]	89.10 [88.06, 90.14]
Pseudo-linear	56.52 [55.52, 57.49]	68.92 [68.00, 69.85]	−1.98 [−2.28, −1.68]	+1.87 [+1.58, +2.16]	56.57 [55.58, 57.54]	48.55 [47.39, 49.74]	86.06 [84.96, 87.14]
ZLaP	32.30 [31.42, 33.18]	70.54 [69.60, 71.48]	−0.35 [−0.79, +0.07]	−23.97 [−24.80, −23.15]	53.85 [52.77, 54.93]	21.60 [21.09, 22.13]	78.34 [77.15, 79.54]
GDA-CLIP	34.23 [33.41, 35.05]	67.27 [66.41, 68.15]	−3.63 [−4.05, −3.22]	−18.78 [−19.47, −18.03]	56.25 [55.23, 57.27]	19.41 [19.03, 19.80]	87.31 [86.15, 88.42]
TransCLIP	37.21 [36.43, 38.02]	66.45 [65.46, 67.44]	−4.44 [−4.95, −3.93]	−14.97 [−15.66, −14.26]	56.62 [55.64, 57.59]	21.08 [20.73, 21.46]	97.36 [96.83, 97.86]
ADAPT	56.35 [55.34, 57.36]	67.94 [67.03, 68.84]	−2.96 [−3.39, −2.54]	+2.69 [+2.29, +3.09]	56.53 [55.54, 57.51]	48.83 [47.66, 50.02]	86.21 [85.09, 87.31]
StatA	60.12 [59.07, 61.17]	67.72 [66.70, 68.76]	−3.17 [−3.66, −2.69]	+6.67 [+6.19, +7.14]	60.19 [59.21, 61.17]	63.75 [62.60, 64.90]	81.41 [80.37, 82.46]
MOON	60.64 [59.56, 61.72]	65.66 [64.70, 66.63]	−5.24 [−5.73, −4.76]	+9.25 [+8.75, +9.75]	60.46 [59.40, 61.52]	72.32 [70.87, 73.76]	73.41 [72.22, 74.59]
MOON (soft)	60.73 [59.65, 61.80]	65.19 [64.24, 66.15]	−5.71 [−6.23, −5.20]	+9.81 [+9.30, +10.33]	60.67 [59.60, 61.72]	77.66 [76.09, 79.20]	73.97 [72.65, 75.25]

Table 27: OpenAI ViT-B/32 matched random context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	62.04 [61.18, 62.88]	87.80 [87.17, 88.44]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.04 [61.18, 62.88]	40.78 [39.85, 41.70]	88.20 [87.31, 89.06]
Pseudo-linear	62.15 [61.28, 63.00]	85.11 [84.41, 85.79]	−2.69 [−2.99, −2.39]	+2.79 [+2.50, +3.10]	62.13 [61.27, 62.98]	40.88 [39.95, 41.81]	86.95 [86.04, 87.84]
ZLaP	36.00 [35.25, 36.76]	86.20 [85.47, 86.92]	−1.59 [−1.95, −1.24]	−24.44 [−25.23, −23.62]	59.93 [59.01, 60.85]	19.54 [19.13, 19.95]	79.17 [78.08, 80.25]
GDA-CLIP	33.34 [32.76, 33.92]	80.42 [79.82, 81.03]	−7.37 [−7.75, −7.01]	−21.33 [−22.13, −20.51]	61.20 [60.31, 62.08]	16.93 [16.68, 17.19]	84.49 [83.50, 85.45]
TransCLIP	36.63 [36.06, 37.19]	81.66 [80.83, 82.48]	−6.14 [−6.76, −5.53]	−19.27 [−20.08, −18.45]	61.99 [61.13, 62.83]	18.13 [17.91, 18.34]	93.79 [93.15, 94.44]
ADAPT	63.00 [62.10, 63.89]	81.67 [80.84, 82.48]	−6.13 [−6.70, −5.56]	+7.09 [+6.51, +7.67]	62.70 [61.82, 63.58]	42.18 [41.24, 43.11]	86.60 [85.65, 87.54]
StatA	67.70 [66.76, 68.61]	84.33 [83.65, 85.03]	−3.46 [−3.89, −3.05]	+9.12 [+8.58, +9.68]	67.35 [66.42, 68.26]	54.81 [53.66, 55.96]	81.64 [80.63, 82.63]
MOON	69.81 [68.76, 70.82]	80.81 [79.97, 81.64]	−6.99 [−7.55, −6.42]	+14.76 [+14.06, +15.46]	69.39 [68.34, 70.41]	66.84 [65.57, 68.04]	76.34 [75.28, 77.37]
MOON (soft)	70.28 [69.21, 71.30]	80.79 [79.95, 81.61]	−7.01 [−7.57, −6.46]	+15.25 [+14.57, +15.94]	70.13 [69.07, 71.16]	72.25 [71.12, 73.34]	75.72 [74.63, 76.79]

Table 28: OpenAI ViT-B/32 matched far context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	73.41 [72.71, 74.10]	90.78 [90.20, 91.35]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	73.41 [72.71, 74.10]	49.44 [48.44, 50.47]	94.60 [94.03, 95.15]
Pseudo-linear	73.64 [72.94, 74.33]	89.89 [89.27, 90.51]	−0.89 [−1.10, −0.69]	+1.12 [+0.91, +1.34]	73.59 [72.89, 74.28]	49.87 [48.88, 50.89]	94.17 [93.57, 94.77]
ZLaP	43.83 [43.09, 44.57]	90.31 [89.66, 90.96]	−0.47 [−0.76, −0.17]	−29.11 [−29.78, −28.43]	72.23 [71.45, 73.03]	22.74 [22.34, 23.15]	88.63 [87.85, 89.41]
GDA-CLIP	41.28 [40.63, 41.92]	82.72 [82.15, 83.29]	−8.06 [−8.42, −7.70]	−24.07 [−24.83, −23.30]	73.15 [72.39, 73.90]	19.96 [19.72, 20.19]	94.02 [93.44, 94.58]
TransCLIP	44.24 [43.64, 44.85]	87.06 [86.36, 87.76]	−3.72 [−4.13, −3.32]	−25.45 [−26.11, −24.80]	73.41 [72.71, 74.10]	20.37 [20.15, 20.59]	97.63 [97.26, 98.00]
ADAPT	75.45 [74.74, 76.16]	88.37 [87.68, 89.04]	−2.41 [−2.80, −2.05]	+4.45 [+4.06, +4.85]	75.03 [74.33, 75.72]	53.67 [52.56, 54.81]	93.78 [93.11, 94.42]
StatA	76.83 [76.02, 77.62]	87.88 [87.18, 88.56]	−2.90 [−3.25, −2.54]	+6.31 [+5.88, +6.74]	77.49 [76.74, 78.22]	61.55 [60.46, 62.65]	90.18 [89.38, 90.96]
MOON	81.57 [80.82, 82.32]	87.55 [86.89, 88.20]	−3.23 [−3.56, −2.91]	+11.39 [+10.93, +11.85]	81.11 [80.33, 81.88]	78.71 [77.66, 79.76]	88.77 [87.97, 89.56]
MOON (soft)	81.88 [81.13, 82.62]	87.07 [86.39, 87.72]	−3.71 [−4.05, −3.39]	+12.18 [+11.68, +12.67]	81.68 [80.90, 82.47]	83.21 [82.26, 84.16]	88.46 [87.74, 89.17]

Table 29: LAION ViT-B/32 matched near context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.55 [63.24, 67.80]	72.68 [70.50, 74.80]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.55 [63.24, 67.80]	51.94 [50.86, 53.05]	90.43 [89.49, 91.35]
Pseudo-linear	65.52 [63.21, 67.79]	72.36 [70.11, 74.57]	−0.32 [−0.83, +0.16]	+0.30 [−0.17, +0.79]	65.52 [63.21, 67.78]	51.79 [50.73, 52.90]	89.51 [88.58, 90.44]
ZLaP	46.45 [43.91, 48.88]	72.26 [70.10, 74.38]	−0.42 [−1.90, +1.06]	−18.68 [−20.73, −16.74]	64.74 [62.13, 67.22]	25.93 [25.36, 26.50]	85.74 [84.76, 86.75]
GDA-CLIP	48.08 [45.91, 50.28]	67.70 [65.73, 69.66]	−4.98 [−6.33, −3.62]	−12.48 [−14.11, −10.84]	65.91 [63.58, 68.21]	21.46 [21.08, 21.82]	91.67 [90.83, 92.47]
TransCLIP	50.18 [48.02, 52.37]	71.71 [69.54, 73.86]	−0.97 [−2.45, +0.54]	−14.40 [−16.07, −12.78]	65.55 [63.24, 67.80]	23.48 [23.10, 23.86]	98.15 [97.68, 98.59]
ADAPT	65.95 [63.65, 68.20]	72.94 [70.60, 75.24]	+0.26 [−0.77, +1.22]	+0.14 [−0.59, +0.96]	65.59 [63.28, 67.83]	52.02 [50.95, 53.12]	89.74 [88.77, 90.70]
StatA	67.40 [64.97, 69.80]	71.55 [69.18, 73.88]	−1.13 [−2.37, +0.12]	+2.99 [+2.01, +3.94]	67.54 [65.22, 69.83]	63.55 [62.37, 64.75]	85.32 [84.34, 86.30]
MOON	68.84 [66.40, 71.23]	72.11 [69.66, 74.54]	−0.57 [−1.59, +0.41]	+3.86 [+3.03, +4.72]	68.16 [65.81, 70.52]	70.84 [69.49, 72.23]	80.82 [79.70, 81.89]
MOON (soft)	69.12 [66.69, 71.52]	71.80 [69.38, 74.18]	−0.88 [−1.92, +0.10]	+4.45 [+3.58, +5.36]	68.82 [66.45, 71.15]	77.33 [76.03, 78.65]	81.28 [80.25, 82.30]

Table 30: LAION ViT-B/32 matched random context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.55 [63.24, 67.80]	89.11 [87.42, 90.75]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.55 [63.24, 67.80]	43.67 [42.83, 44.57]	89.80 [89.00, 90.58]
Pseudo-linear	65.52 [63.21, 67.77]	88.16 [86.31, 90.00]	−0.95 [−1.78, −0.17]	+0.93 [+0.15, +1.76]	65.55 [63.24, 67.80]	43.66 [42.82, 44.57]	89.48 [88.67, 90.26]
ZLaP	42.08 [39.81, 44.28]	88.81 [87.06, 90.48]	−0.30 [−1.18, +0.54]	−23.17 [−25.35, −21.04]	64.35 [61.68, 66.90]	22.36 [21.95, 22.77]	84.62 [83.69, 85.55]
GDA-CLIP	37.52 [35.76, 39.29]	78.74 [76.89, 80.56]	−10.37 [−11.84, −8.97]	−17.66 [−19.86, −15.38]	65.27 [62.85, 67.62]	18.08 [17.88, 18.29]	88.74 [87.94, 89.52]
TransCLIP	40.87 [39.02, 42.73]	85.47 [83.62, 87.33]	−3.63 [−5.08, −2.20]	−21.05 [−23.15, −18.95]	65.57 [63.27, 67.80]	19.58 [19.37, 19.80]	95.92 [95.37, 96.44]
ADAPT	66.04 [63.71, 68.31]	85.68 [83.58, 87.81]	−3.43 [−4.84, −2.09]	+3.92 [+2.55, +5.35]	65.89 [63.57, 68.16]	44.07 [43.23, 44.99]	89.45 [88.65, 90.23]
StatA	70.84 [68.41, 73.28]	86.57 [84.60, 88.53]	−2.54 [−3.93, −1.21]	+7.83 [+6.26, +9.44]	70.32 [67.97, 72.67]	56.32 [55.23, 57.45]	84.97 [83.99, 85.94]
MOON	73.05 [70.63, 75.52]	86.30 [84.20, 88.34]	−2.81 [−3.99, −1.67]	+10.32 [+8.98, +11.68]	71.92 [69.50, 74.35]	64.77 [63.67, 65.89]	82.84 [81.95, 83.73]
MOON (soft)	75.05 [72.63, 77.48]	86.42 [84.31, 88.50]	−2.68 [−3.84, −1.56]	+12.19 [+10.58, +13.80]	74.34 [71.94, 76.72]	71.30 [70.20, 72.42]	81.77 [80.83, 82.72]

Table 31: LAION ViT-B/32 matched far context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	65.55 [63.24, 67.80]	97.35 [96.34, 98.24]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.55 [63.24, 67.80]	49.26 [48.32, 50.23]	93.66 [93.05, 94.26]
Pseudo-linear	65.52 [63.21, 67.77]	97.06 [95.93, 98.05]	−0.29 [−0.65, +0.06]	+0.27 [−0.09, +0.62]	65.52 [63.21, 67.77]	49.23 [48.30, 50.19]	93.22 [92.56, 93.86]
ZLaP	38.82 [36.78, 40.88]	94.75 [93.19, 96.24]	−2.60 [−3.83, −1.48]	−24.13 [−26.61, −21.64]	63.41 [60.68, 66.04]	25.31 [24.84, 25.82]	91.48 [90.77, 92.16]
GDA-CLIP	29.62 [27.95, 31.32]	79.92 [78.16, 81.61]	−17.44 [−18.97, −15.89]	−18.49 [−21.31, −15.71]	62.76 [60.16, 65.29]	20.44 [20.16, 20.73]	94.49 [93.89, 95.08]
TransCLIP	33.97 [32.36, 35.64]	96.40 [95.30, 97.39]	−0.95 [−1.79, −0.10]	−30.63 [−32.72, −28.53]	65.55 [63.24, 67.80]	21.50 [21.21, 21.79]	97.62 [97.16, 98.06]
ADAPT	65.88 [63.56, 68.16]	94.18 [92.51, 95.72]	−3.17 [−4.37, −2.08]	+3.50 [+2.36, +4.74]	65.80 [63.48, 68.09]	49.90 [48.92, 50.93]	93.42 [92.79, 94.03]
StatA	73.12 [70.73, 75.52]	96.53 [95.38, 97.57]	−0.82 [−1.56, −0.12]	+8.39 [+6.90, +9.95]	72.63 [70.22, 74.98]	61.20 [60.12, 62.27]	89.97 [89.25, 90.70]
MOON	75.36 [72.82, 77.92]	95.89 [94.53, 97.10]	−1.47 [−2.11, −0.88]	+11.28 [+9.80, +12.79]	74.57 [72.01, 77.11]	71.01 [69.94, 72.08]	88.83 [88.11, 89.55]
MOON (soft)	76.28 [73.81, 78.81]	95.80 [94.39, 97.05]	−1.55 [−2.24, −0.94]	+12.29 [+10.62, +14.07]	75.79 [73.31, 78.30]	77.25 [76.31, 78.19]	88.73 [88.08, 89.38]

Table 32: LAION ViT-B/32 matched near context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	62.86 [61.95, 63.79]	76.04 [75.21, 76.88]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.86 [61.95, 63.79]	51.94 [50.86, 53.05]	90.43 [89.49, 91.35]
Pseudo-linear	62.89 [61.97, 63.81]	75.15 [74.30, 76.02]	−0.89 [−1.11, −0.68]	+0.92 [+0.71, +1.13]	62.88 [61.97, 63.79]	51.79 [50.73, 52.90]	89.51 [88.58, 90.44]
ZLaP	41.80 [40.98, 42.63]	75.96 [75.11, 76.81]	−0.08 [−0.50, +0.33]	−20.98 [−21.70, −20.23]	61.30 [60.32, 62.27]	25.93 [25.36, 26.50]	85.74 [84.76, 86.75]
GDA-CLIP	40.24 [39.48, 40.99]	69.43 [68.64, 70.22]	−6.61 [−7.05, −6.19]	−16.02 [−16.79, −15.24]	62.88 [61.98, 63.78]	21.46 [21.08, 21.82]	91.67 [90.83, 92.47]
TransCLIP	43.51 [42.75, 44.29]	73.22 [72.31, 74.12]	−2.83 [−3.36, −2.29]	−16.53 [−17.22, −15.83]	62.87 [61.96, 63.80]	23.48 [23.10, 23.86]	98.15 [97.68, 98.59]
ADAPT	62.92 [62.00, 63.87]	74.10 [73.23, 74.98]	−1.94 [−2.34, −1.55]	+2.00 [+1.64, +2.37]	62.88 [61.96, 63.80]	52.02 [50.95, 53.12]	89.74 [88.77, 90.70]
StatA	66.46 [65.49, 67.47]	74.44 [73.52, 75.37]	−1.60 [−1.99, −1.22]	+5.20 [+4.81, +5.60]	66.29 [65.35, 67.27]	63.55 [62.37, 64.75]	85.32 [84.34, 86.30]
MOON	67.58 [66.57, 68.61]	73.49 [72.57, 74.42]	−2.56 [−2.93, −2.20]	+7.27 [+6.85, +7.69]	66.96 [65.97, 67.97]	70.84 [69.49, 72.23]	80.82 [79.70, 81.89]
MOON (soft)	68.39 [67.38, 69.42]	73.29 [72.39, 74.19]	−2.76 [−3.13, −2.39]	+8.28 [+7.82, +8.72]	68.01 [67.01, 69.02]	77.33 [76.03, 78.65]	81.28 [80.25, 82.30]

Table 33: LAION ViT-B/32 matched random context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	66.14 [65.40, 66.90]	89.58 [88.99, 90.15]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.14 [65.40, 66.90]	43.67 [42.83, 44.57]	89.80 [89.00, 90.58]
Pseudo-linear	66.16 [65.41, 66.93]	88.31 [87.68, 88.93]	−1.26 [−1.49, −1.04]	+1.28 [+1.06, +1.51]	66.16 [65.40, 66.92]	43.66 [42.82, 44.57]	89.48 [88.67, 90.26]
ZLaP	43.16 [42.40, 43.90]	88.76 [88.12, 89.38]	−0.82 [−1.17, −0.48]	−22.16 [−22.96, −21.39]	65.00 [64.13, 65.88]	22.36 [21.95, 22.77]	84.62 [83.69, 85.55]
GDA-CLIP	37.00 [36.46, 37.55]	79.48 [78.87, 80.09]	−10.10 [−10.50, −9.70]	−19.05 [−19.84, −18.29]	66.07 [65.28, 66.86]	18.08 [17.88, 18.29]	88.74 [87.94, 89.52]
TransCLIP	40.53 [39.98, 41.08]	85.82 [85.16, 86.48]	−3.76 [−4.23, −3.29]	−21.86 [−22.59, −21.14]	66.17 [65.41, 66.93]	19.58 [19.37, 19.80]	95.92 [95.37, 96.44]
ADAPT	66.54 [65.78, 67.31]	85.57 [84.87, 86.24]	−4.01 [−4.45, −3.58]	+4.41 [+3.99, +4.84]	66.35 [65.59, 67.11]	44.07 [43.23, 44.99]	89.45 [88.65, 90.23]
StatA	71.84 [70.96, 72.70]	87.56 [86.93, 88.17]	−2.02 [−2.38, −1.66]	+7.71 [+7.19, +8.21]	71.28 [70.41, 72.14]	56.32 [55.23, 57.45]	84.97 [83.99, 85.94]
MOON	73.81 [72.95, 74.69]	86.26 [85.59, 86.92]	−3.32 [−3.67, −2.97]	+10.98 [+10.47, +11.49]	72.97 [72.13, 73.84]	64.77 [63.67, 65.89]	82.84 [81.95, 83.73]
MOON (soft)	74.92 [74.03, 75.84]	86.28 [85.61, 86.95]	−3.29 [−3.64, −2.96]	+12.07 [+11.50, +12.64]	74.43 [73.54, 75.33]	71.30 [70.20, 72.42]	81.77 [80.83, 82.72]

Table 34: LAION ViT-B/32 matched far context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	73.47 [72.80, 74.16]	91.23 [90.63, 91.81]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	73.47 [72.80, 74.16]	49.26 [48.32, 50.23]	93.66 [93.05, 94.26]
Pseudo-linear	73.53 [72.85, 74.21]	90.34 [89.67, 90.96]	−0.89 [−1.06, −0.73]	+0.94 [+0.78, +1.11]	73.49 [72.81, 74.17]	49.23 [48.30, 50.19]	93.22 [92.56, 93.86]
ZLaP	50.30 [49.57, 51.04]	90.98 [90.34, 91.61]	−0.24 [−0.57, +0.07]	−22.93 [−23.63, −22.22]	73.91 [73.15, 74.65]	25.31 [24.84, 25.82]	91.48 [90.77, 92.16]
GDA-CLIP	43.49 [42.84, 44.12]	81.26 [80.67, 81.85]	−9.97 [−10.35, −9.58]	−20.01 [−20.79, −19.26]	73.51 [72.79, 74.21]	20.44 [20.16, 20.73]	94.49 [93.89, 95.08]
TransCLIP	47.11 [46.47, 47.73]	89.13 [88.48, 89.77]	−2.10 [−2.46, −1.73]	−24.27 [−24.91, −23.65]	73.46 [72.79, 74.15]	21.50 [21.21, 21.79]	97.62 [97.16, 98.06]
ADAPT	73.97 [73.30, 74.65]	88.52 [87.85, 89.17]	−2.71 [−3.06, −2.36]	+3.21 [+2.87, +3.56]	73.77 [73.10, 74.45]	49.90 [48.92, 50.93]	93.42 [92.79, 94.03]
StatA	77.68 [76.90, 78.44]	89.40 [88.67, 90.10]	−1.83 [−2.19, −1.48]	+6.03 [+5.62, +6.45]	77.50 [76.77, 78.23]	61.20 [60.12, 62.27]	89.97 [89.25, 90.70]
MOON	80.42 [79.67, 81.16]	88.78 [88.12, 89.42]	−2.45 [−2.72, −2.18]	+9.39 [+8.99, +9.79]	79.54 [78.81, 80.28]	71.01 [69.94, 72.08]	88.83 [88.11, 89.55]
MOON (soft)	81.28 [80.54, 82.01]	88.55 [87.89, 89.19]	−2.68 [−2.97, −2.39]	+10.48 [+10.05, +10.93]	80.56 [79.82, 81.28]	77.25 [76.31, 78.19]	88.73 [88.08, 89.38]

Table 35: OpenAI ViT-B/16 matched near context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	63.44 [61.04, 65.83]	69.21 [66.83, 71.58]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	63.44 [61.04, 65.83]	53.81 [52.57, 55.07]	87.55 [86.45, 88.61]
Pseudo-linear	63.70 [61.25, 66.15]	69.05 [66.55, 71.57]	−0.16 [−0.81, +0.47]	+0.42 [−0.15, +1.01]	63.51 [61.08, 65.92]	53.75 [52.51, 55.01]	85.58 [84.48, 86.66]
ZLaP	43.23 [40.65, 45.71]	69.94 [67.58, 72.29]	+0.73 [−0.75, +2.23]	−20.94 [−22.81, −19.10]	62.75 [60.00, 65.48]	24.17 [23.53, 24.82]	82.07 [80.97, 83.14]
GDA-CLIP	45.84 [43.55, 48.16]	68.57 [66.54, 70.61]	−0.64 [−2.02, +0.80]	−16.96 [−18.57, −15.34]	63.84 [61.42, 66.26]	20.38 [20.03, 20.74]	89.53 [88.55, 90.48]
TransCLIP	47.30 [44.91, 49.70]	68.61 [65.99, 71.12]	−0.60 [−2.25, +1.03]	−15.54 [−17.25, −13.88]	63.46 [61.05, 65.84]	22.03 [21.66, 22.40]	97.26 [96.67, 97.81]
ADAPT	63.76 [61.32, 66.21]	68.73 [66.18, 71.26]	−0.48 [−1.47, +0.49]	+0.79 [+0.18, +1.41]	63.69 [61.26, 66.09]	54.38 [53.14, 55.65]	85.88 [84.84, 86.93]
StatA	64.16 [61.36, 66.92]	67.24 [64.42, 70.00]	−1.97 [−3.48, −0.50]	+2.68 [+1.77, +3.63]	64.78 [62.10, 67.44]	68.64 [67.45, 69.87]	80.38 [79.26, 81.47]
MOON	65.68 [62.85, 68.42]	67.64 [64.72, 70.46]	−1.57 [−2.93, −0.33]	+3.81 [+2.91, +4.77]	65.33 [62.56, 68.04]	75.81 [74.47, 77.15]	74.69 [73.62, 75.77]
MOON (soft)	65.35 [62.53, 68.14]	66.65 [63.83, 69.45]	−2.56 [−3.82, −1.38]	+4.47 [+3.55, +5.43]	65.66 [63.00, 68.29]	81.21 [79.82, 82.54]	76.44 [75.29, 77.59]

Table 36: OpenAI ViT-B/16 matched random context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	63.44 [61.04, 65.83]	88.85 [87.05, 90.63]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	63.44 [61.04, 65.83]	44.47 [43.60, 45.38]	87.52 [86.65, 88.39]
Pseudo-linear	63.80 [61.36, 66.22]	86.71 [84.63, 88.77]	−2.15 [−3.34, −1.03]	+2.51 [+1.39, +3.69]	63.75 [61.33, 66.16]	44.76 [43.87, 45.68]	86.76 [85.85, 87.66]
ZLaP	38.26 [35.91, 40.65]	88.47 [86.61, 90.34]	−0.39 [−1.59, +0.79]	−24.80 [−27.24, −22.31]	62.01 [59.21, 64.81]	21.38 [20.93, 21.87]	81.53 [80.49, 82.58]
GDA-CLIP	34.77 [32.81, 36.68]	82.39 [80.64, 84.15]	−6.47 [−7.80, −5.16]	−22.21 [−24.36, −20.08]	63.62 [61.14, 66.10]	17.78 [17.55, 18.01]	86.39 [85.47, 87.28]
TransCLIP	37.38 [35.36, 39.41]	85.06 [82.90, 87.18]	−3.79 [−5.66, −2.00]	−22.27 [−24.77, −19.81]	63.57 [61.16, 65.95]	18.94 [18.72, 19.17]	93.77 [93.12, 94.41]
ADAPT	64.69 [62.26, 67.15]	82.66 [80.14, 85.12]	−6.20 [−8.05, −4.45]	+7.44 [+5.69, +9.29]	64.39 [61.95, 66.85]	46.23 [45.33, 47.20]	86.34 [85.46, 87.22]
StatA	69.53 [66.69, 72.26]	84.34 [81.93, 86.63]	−4.51 [−6.30, −2.92]	+10.60 [+8.74, +12.48]	69.72 [67.05, 72.31]	59.18 [58.07, 60.32]	80.41 [79.34, 81.47]
MOON	71.55 [68.54, 74.50]	82.35 [79.63, 84.96]	−6.50 [−8.44, −4.76]	+14.61 [+12.70, +16.61]	71.21 [68.20, 74.10]	69.82 [68.65, 70.98]	77.53 [76.45, 78.60]
MOON (soft)	72.04 [69.18, 74.83]	81.67 [79.03, 84.22]	−7.18 [−8.99, −5.53]	+15.78 [+13.87, +17.73]	72.11 [69.22, 74.89]	73.64 [72.51, 74.72]	76.64 [75.54, 77.71]

Table 37: OpenAI ViT-B/16 matched far context, evaluated on $Q = 8$ fixed target queries; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	63.44 [61.04, 65.83]	97.63 [96.53, 98.61]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	63.44 [61.04, 65.83]	51.85 [50.88, 52.83]	93.88 [93.31, 94.44]
Pseudo-linear	63.72 [61.28, 66.14]	96.75 [95.44, 97.93]	−0.88 [−1.42, −0.36]	+1.15 [+0.60, +1.71]	63.58 [61.16, 65.99]	52.40 [51.43, 53.41]	93.39 [92.77, 94.00]
ZLaP	34.42 [32.42, 36.41]	95.47 [93.86, 96.99]	−2.15 [−3.38, −1.05]	−26.87 [−29.20, −24.51]	62.12 [59.31, 64.94]	24.54 [24.15, 24.92]	88.92 [88.12, 89.72]
GDA-CLIP	26.70 [24.88, 28.53]	84.61 [83.11, 86.04]	−13.02 [−14.38, −11.64]	−23.73 [−26.31, −21.13]	61.90 [59.19, 64.65]	20.67 [20.39, 20.95]	93.28 [92.67, 93.88]
TransCLIP	31.15 [29.42, 32.90]	95.95 [94.64, 97.15]	−1.68 [−2.75, −0.70]	−30.62 [−32.92, −28.29]	63.69 [61.27, 66.10]	21.52 [21.21, 21.84]	97.52 [97.05, 97.97]
ADAPT	64.98 [62.52, 67.45]	91.65 [89.54, 93.65]	−5.98 [−7.74, −4.31]	+7.52 [+5.81, +9.32]	64.51 [62.04, 66.96]	55.02 [53.97, 56.09]	92.97 [92.29, 93.62]
StatA	71.20 [68.33, 73.94]	96.02 [94.58, 97.33]	−1.60 [−2.60, −0.72]	+9.36 [+7.53, +11.16]	71.30 [68.55, 73.99]	66.57 [65.41, 67.69]	87.69 [86.75, 88.59]
MOON	74.16 [71.19, 77.04]	93.96 [92.06, 95.68]	−3.67 [−5.14, −2.40]	+14.38 [+12.41, +16.39]	73.80 [70.83, 76.65]	80.22 [79.13, 81.28]	87.40 [86.50, 88.28]
MOON (soft)	74.95 [72.04, 77.73]	93.96 [92.04, 95.69]	−3.67 [−5.14, −2.40]	+15.17 [+13.20, +17.23]	74.80 [71.90, 77.59]	83.50 [82.49, 84.46]	87.45 [86.59, 88.28]

Table 38: OpenAI ViT-B/16 matched near context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	60.66 [59.64, 61.66]	72.89 [71.90, 73.87]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	60.66 [59.64, 61.66]	53.81 [52.57, 55.07]	87.55 [86.45, 88.61]
Pseudo-linear	60.71 [59.70, 61.72]	71.23 [70.25, 72.20]	−1.66 [−1.93, −1.40]	+1.71 [+1.46, +1.97]	60.67 [59.67, 61.69]	53.75 [52.51, 55.01]	85.58 [84.48, 86.66]
ZLaP	37.09 [36.09, 38.09]	73.69 [72.77, 74.60]	+0.79 [+0.36, +1.22]	−24.36 [−25.23, −23.50]	58.75 [57.66, 59.83]	24.17 [23.53, 24.82]	82.07 [80.97, 83.14]
GDA-CLIP	37.14 [36.38, 37.91]	70.86 [69.94, 71.77]	−2.03 [−2.48, −1.57]	−21.48 [−22.23, −20.73]	60.78 [59.76, 61.81]	20.38 [20.03, 20.74]	89.53 [88.55, 90.48]
TransCLIP	40.78 [39.98, 41.60]	70.85 [69.84, 71.85]	−2.05 [−2.64, −1.47]	−17.83 [−18.61, −17.06]	60.66 [59.66, 61.67]	22.03 [21.66, 22.40]	97.26 [96.67, 97.81]
ADAPT	61.03 [60.05, 62.05]	69.99 [69.01, 70.97]	−2.90 [−3.39, −2.42]	+3.28 [+2.85, +3.71]	60.80 [59.80, 61.80]	54.38 [53.14, 55.65]	85.88 [84.84, 86.93]
StatA	63.73 [62.67, 64.79]	70.40 [69.32, 71.48]	−2.50 [−2.96, −2.03]	+5.57 [+5.12, +6.02]	63.74 [62.72, 64.75]	68.64 [67.45, 69.87]	80.38 [79.26, 81.47]
MOON	63.98 [62.94, 65.02]	67.80 [66.81, 68.79]	−5.09 [−5.56, −4.64]	+8.42 [+7.91, +8.92]	63.72 [62.70, 64.76]	75.81 [74.47, 77.15]	74.69 [73.62, 75.77]
MOON (soft)	64.42 [63.37, 65.48]	67.60 [66.60, 68.64]	−5.29 [−5.79, −4.80]	+9.05 [+8.55, +9.57]	64.39 [63.36, 65.44]	81.21 [79.82, 82.54]	76.44 [75.29, 77.59]

Table 39: OpenAI ViT-B/16 matched random context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	64.72 [63.91, 65.52]	89.70 [89.06, 90.33]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	64.72 [63.91, 65.52]	44.47 [43.60, 45.38]	87.52 [86.65, 88.39]
Pseudo-linear	64.94 [64.11, 65.76]	87.11 [86.42, 87.79]	−2.58 [−2.89, −2.28]	+2.81 [+2.51, +3.12]	64.92 [64.10, 65.74]	44.76 [43.87, 45.68]	86.76 [85.85, 87.66]
ZLaP	40.38 [39.62, 41.12]	88.92 [88.20, 89.61]	−0.78 [−1.11, −0.44]	−23.56 [−24.33, −22.84]	63.75 [62.80, 64.70]	21.38 [20.93, 21.87]	81.53 [80.49, 82.58]
GDA-CLIP	35.94 [35.35, 36.52]	83.67 [83.07, 84.26]	−6.02 [−6.46, −5.60]	−22.75 [−23.52, −21.97]	64.57 [63.71, 65.42]	17.78 [17.55, 18.01]	86.39 [85.47, 87.28]
TransCLIP	39.03 [38.44, 39.60]	85.67 [84.89, 86.41]	−4.03 [−4.63, −3.43]	−21.66 [−22.45, −20.88]	64.74 [63.93, 65.55]	18.94 [18.72, 19.17]	93.77 [93.12, 94.41]
ADAPT	65.62 [64.77, 66.44]	82.75 [81.95, 83.55]	−6.94 [−7.54, −6.34]	+7.84 [+7.23, +8.46]	65.43 [64.62, 66.25]	46.23 [45.33, 47.20]	86.34 [85.46, 87.22]
StatA	69.92 [68.95, 70.87]	86.72 [85.97, 87.47]	−2.97 [−3.39, −2.57]	+8.17 [+7.61, +8.75]	69.60 [68.66, 70.54]	59.18 [58.07, 60.32]	80.41 [79.34, 81.47]
MOON	71.88 [70.87, 72.87]	83.19 [82.37, 84.02]	−6.51 [−7.03, −5.99]	+13.67 [+12.97, +14.36]	71.49 [70.49, 72.47]	69.82 [68.65, 70.98]	77.53 [76.45, 78.60]
MOON (soft)	72.05 [71.04, 73.04]	83.07 [82.22, 83.92]	−6.63 [−7.18, −6.10]	+13.96 [+13.25, +14.68]	71.94 [70.91, 72.93]	73.64 [72.51, 74.72]	76.64 [75.54, 77.71]

Table 40: OpenAI ViT-B/16 matched far context, evaluated on all $N = 64$ images; five-dataset macro, 540 target-class episodes. Mean [95% CI]. All terms except P and R use the specified evaluation set; P and R use the full episode.

Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
Frozen CLIP	74.12 [73.31, 74.88]	91.97 [91.24, 92.68]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	74.12 [73.31, 74.88]	51.85 [50.88, 52.83]	93.88 [93.31, 94.44]
Pseudo-linear	74.46 [73.66, 75.23]	91.04 [90.31, 91.76]	−0.92 [−1.11, −0.74]	+1.27 [+1.07, +1.46]	74.30 [73.50, 75.07]	52.40 [51.43, 53.41]	93.39 [92.77, 94.00]
ZLaP	48.03 [47.31, 48.72]	91.98 [91.27, 92.68]	+0.01 [−0.27, +0.29]	−26.10 [−26.78, −25.44]	74.44 [73.58, 75.28]	24.54 [24.15, 24.92]	88.92 [88.12, 89.72]
GDA-CLIP	43.42 [42.72, 44.09]	85.45 [84.79, 86.11]	−6.52 [−6.87, −6.16]	−24.18 [−24.92, −23.44]	74.08 [73.27, 74.88]	20.67 [20.39, 20.95]	93.28 [92.67, 93.88]
TransCLIP	46.64 [45.93, 47.34]	89.64 [88.88, 90.39]	−2.32 [−2.70, −1.96]	−25.16 [−25.79, −24.51]	74.16 [73.36, 74.93]	21.52 [21.21, 21.84]	97.52 [97.05, 97.97]
ADAPT	75.55 [74.72, 76.35]	88.65 [87.87, 89.41]	−3.32 [−3.74, −2.91]	+4.75 [+4.33, +5.18]	75.09 [74.26, 75.88]	55.02 [53.97, 56.09]	92.97 [92.29, 93.62]
StatA	77.52 [76.57, 78.42]	89.88 [89.13, 90.61]	−2.09 [−2.43, −1.75]	+5.49 [+5.04, +5.94]	77.68 [76.75, 78.57]	66.57 [65.41, 67.69]	87.69 [86.75, 88.59]
MOON	81.36 [80.40, 82.29]	88.75 [87.93, 89.55]	−3.22 [−3.60, −2.86]	+10.46 [+9.98, +10.94]	81.05 [80.11, 81.95]	80.22 [79.13, 81.28]	87.40 [86.50, 88.28]
MOON (soft)	81.58 [80.63, 82.50]	88.59 [87.76, 89.39]	−3.38 [−3.76, −3.02]	+10.84 [+10.34, +11.36]	81.44 [80.49, 82.37]	83.50 [82.49, 84.46]	87.45 [86.59, 88.28]

D.2 PAIRED CONTEXT CONTRASTS

These tables compare the gain over the condition-specific frozen baseline on aligned target classes. Full-gain contrast equals rank contrast plus relief contrast. Pairing removes variation due to the fixed query selection when the contexts are compared.

Table 41: Paired matched-context contrast (near minus far), OpenAI ViT-B/32, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.18\,[-0.44, +0.06]$	$+1.02\,[+0.13, +1.95]$	$-1.20\,[-2.15, -0.30]$
ZLaP	$+8.06\,[+6.21, +9.98]$	$+1.56\,[-0.24, +3.33]$	$+6.50\,[+4.06, +8.97]$
GDA-CLIP	$+18.83\,[+17.11, +20.61]$	$+12.45\,[+10.67, +14.28]$	$+6.38\,[+3.70, +9.01]$
TransCLIP	$+15.40\,[+13.64, +17.20]$	$+0.59\,[-1.35, +2.54]$	$+14.82\,[+12.48, +17.16]$
ADAPT	$-2.41\,[-3.28, -1.59]$	$+4.63\,[+2.89, +6.46]$	$-7.04\,[-8.73, -5.44]$
StatA	$-8.02\,[-9.84, -6.30]$	$-0.37\,[-2.23, +1.44]$	$-7.66\,[-9.43, -5.92]$
MOON	$-7.79\,[-9.44, -6.21]$	$+2.83\,[+1.30, +4.38]$	$-10.62\,[-12.52, -8.74]$
MOON (soft)	$-8.84\,[-10.55, -7.13]$	$+1.97\,[+0.51, +3.46]$	$-10.81\,[-12.74, -8.95]$

Table 42: Paired matched-context contrast (near minus random), OpenAI ViT-B/32, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.07\,[-0.37, +0.21]$	$+2.18\,[+1.11, +3.28]$	$-2.25\,[-3.36, -1.18]$
ZLaP	$+4.24\,[+2.37, +6.07]$	$-0.37\,[-1.95, +1.19]$	$+4.62\,[+2.33, +6.89]$
GDA-CLIP	$+11.92\,[+10.30, +13.63]$	$+5.42\,[+3.78, +7.06]$	$+6.50\,[+4.39, +8.66]$
TransCLIP	$+9.22\,[+7.65, +10.84]$	$+5.02\,[+2.90, +7.06]$	$+4.20\,[+1.80, +6.59]$
ADAPT	$-1.89\,[-2.74, -1.09]$	$+4.48\,[+2.59, +6.37]$	$-6.37\,[-8.14, -4.65]$
StatA	$-5.29\,[-6.97, -3.73]$	$+1.98\,[-0.07, +3.97]$	$-7.28\,[-9.10, -5.46]$
MOON	$-5.36\,[-6.91, -3.87]$	$+5.26\,[+3.41, +7.13]$	$-10.63\,[-12.42, -8.85]$
MOON (soft)	$-6.31\,[-7.94, -4.74]$	$+4.65\,[+2.90, +6.41]$	$-10.96\,[-12.82, -9.12]$

Table 43: Paired matched-context contrast (random minus far), OpenAI ViT-B/32, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.11\,[-0.31, +0.06]$	$-1.17\,[-2.20, -0.15]$	$+1.06\,[+0.05, +2.08]$
ZLaP	$+3.82\,[+2.35, +5.35]$	$+1.93\,[+0.77, +3.12]$	$+1.88\,[+0.02, +3.77]$
GDA-CLIP	$+6.91\,[+5.64, +8.26]$	$+7.04\,[+5.54, +8.59]$	$-0.12\,[-2.09, +1.86]$
TransCLIP	$+6.18\,[+4.87, +7.57]$	$-4.44\,[-6.16, -2.65]$	$+10.62\,[+8.48, +12.70]$
ADAPT	$-0.52\,[-0.94, -0.11]$	$+0.15\,[-1.65, +2.01]$	$-0.68\,[-2.59, +1.19]$
StatA	$-2.73\,[-3.87, -1.68]$	$-2.35\,[-3.76, -0.95]$	$-0.38\,[-1.95, +1.18]$
MOON	$-2.43\,[-3.52, -1.42]$	$-2.44\,[-4.11, -0.79]$	$+0.01\,[-1.83, +1.86]$
MOON (soft)	$-2.53\,[-3.62, -1.47]$	$-2.67\,[-4.33, -1.02]$	$+0.14\,[-1.65, +1.97]$

Table 44: Paired matched-context contrast (near minus far), OpenAI ViT-B/32, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.34\,[-0.45, -0.24]$	$-1.08\,[-1.45, -0.73]$	$+0.74\,[+0.39, +1.11]$
ZLaP	$+5.26\,[+4.25, +6.26]$	$+0.11\,[-0.37, +0.58]$	$+5.14\,[+4.12, +6.15]$
GDA-CLIP	$+9.73\,[+8.89, +10.57]$	$+4.43\,[+3.89, +4.98]$	$+5.30\,[+4.30, +6.32]$
TransCLIP	$+9.76\,[+9.01, +10.53]$	$-0.72\,[-1.34, -0.09]$	$+10.48\,[+9.57, +11.38]$
ADAPT	$-2.31\,[-2.57, -2.05]$	$-0.55\,[-1.14, +0.03]$	$-1.76\,[-2.33, -1.18]$
StatA	$+0.08\,[-0.57, +0.72]$	$-0.28\,[-0.89, +0.33]$	$+0.36\,[-0.30, +1.02]$
MOON	$-4.15\,[-4.76, -3.55]$	$-2.01\,[-2.62, -1.39]$	$-2.14\,[-2.82, -1.46]$
MOON (soft)	$-4.36\,[-4.99, -3.74]$	$-1.99\,[-2.64, -1.37]$	$-2.37\,[-3.09, -1.67]$

Table 45: Paired matched-context contrast (near minus random), OpenAI ViT-B/32, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.22\,[-0.33, -0.11]$	$+0.71\,[+0.29, +1.13]$	$-0.93\,[-1.35, -0.52]$
ZLaP	$+1.71\,[+0.68, +2.73]$	$+1.24\,[+0.69, +1.77]$	$+0.47\,[-0.59, +1.54]$
GDA-CLIP	$+6.30\,[+5.38, +7.22]$	$+3.75\,[+3.24, +4.26]$	$+2.55\,[+1.52, +3.59]$
TransCLIP	$+5.99\,[+5.20, +6.79]$	$+1.70\,[+0.95, +2.46]$	$+4.29\,[+3.29, +5.31]$
ADAPT	$-1.23\,[-1.51, -0.97]$	$+3.17\,[+2.48, +3.86]$	$-4.40\,[-5.09, -3.72]$
StatA	$-2.16\,[-2.87, -1.49]$	$+0.29\,[-0.33, +0.92]$	$-2.45\,[-3.17, -1.77]$
MOON	$-3.76\,[-4.39, -3.12]$	$+1.75\,[+1.03, +2.50]$	$-5.51\,[-6.31, -4.71]$
MOON (soft)	$-4.14\,[-4.83, -3.45]$	$+1.31\,[+0.58, +2.04]$	$-5.44\,[-6.25, -4.62]$

Table 46: Paired matched-context contrast (random minus far), OpenAI ViT-B/32, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.12\,[-0.21, -0.03]$	$-1.79\,[-2.16, -1.44]$	$+1.67\,[+1.31, +2.04]$
ZLaP	$+3.55\,[+2.61, +4.50]$	$-1.13\,[-1.58, -0.68]$	$+4.67\,[+3.67, +5.69]$
GDA-CLIP	$+3.43\,[+2.53, +4.34]$	$+0.68\,[+0.17, +1.19]$	$+2.75\,[+1.69, +3.82]$
TransCLIP	$+3.77\,[+2.97, +4.57]$	$-2.42\,[-3.17, -1.67]$	$+6.18\,[+5.15, +7.22]$
ADAPT	$-1.08\,[-1.32, -0.82]$	$-3.71\,[-4.33, -3.09]$	$+2.64\,[+2.01, +3.26]$
StatA	$+2.25\,[+1.66, +2.85]$	$-0.57\,[-1.12, -0.02]$	$+2.81\,[+2.13, +3.50]$
MOON	$-0.39\,[-1.01, +0.24]$	$-3.76\,[-4.41, -3.10]$	$+3.37\,[+2.57, +4.20]$
MOON (soft)	$-0.22\,[-0.91, +0.46]$	$-3.30\,[-3.97, -2.63]$	$+3.08\,[+2.23, +3.90]$

Table 47: Paired matched-context contrast (near minus far), LAION ViT-B/32, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$+0.00\,[-0.10, +0.10]$	$-0.03\,[-0.57, +0.52]$	$+0.03\,[-0.50, +0.56]$
ZLaP	$+7.64\,[+5.87, +9.36]$	$+2.18\,[+0.32, +4.06]$	$+5.46\,[+3.20, +7.71]$
GDA-CLIP	$+18.46\,[+16.62, +20.36]$	$+12.45\,[+10.57, +14.35]$	$+6.01\,[+3.27, +8.71]$
TransCLIP	$+16.21\,[+14.62, +17.85]$	$-0.03\,[-1.72, +1.73]$	$+16.24\,[+14.18, +18.30]$
ADAPT	$+0.07\,[-0.49, +0.63]$	$+3.43\,[+2.09, +4.81]$	$-3.36\,[-4.61, -2.14]$
StatA	$-5.72\,[-7.02, -4.50]$	$-0.31\,[-1.66, +1.06]$	$-5.41\,[-7.04, -3.82]$
MOON	$-6.52\,[-7.96, -5.19]$	$+0.89\,[-0.17, +1.97]$	$-7.42\,[-8.94, -5.92]$
MOON (soft)	$-7.16\,[-8.89, -5.51]$	$+0.68\,[-0.46, +1.78]$	$-7.84\,[-9.59, -6.15]$

Table 48: Paired matched-context contrast (near minus random), LAION ViT-B/32, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$+0.00\,[-0.10, +0.10]$	$+0.63\,[-0.22, +1.49]$	$-0.63\,[-1.49, +0.22]$
ZLaP	$+4.38\,[+2.75, +5.97]$	$-0.12\,[-1.71, +1.50]$	$+4.49\,[+2.48, +6.53]$
GDA-CLIP	$+10.57\,[+8.95, +12.21]$	$+5.39\,[+3.79, +7.03]$	$+5.18\,[+2.97, +7.38]$
TransCLIP	$+9.31\,[+7.94, +10.69]$	$+2.66\,[+0.86, +4.47]$	$+6.65\,[+4.54, +8.72]$
ADAPT	$-0.09\,[-0.65, +0.49]$	$+3.69\,[+2.26, +5.21]$	$-3.78\,[-5.19, -2.42]$
StatA	$-3.44\,[-4.51, -2.38]$	$+1.41\,[-0.17, +3.04]$	$-4.84\,[-6.40, -3.34]$
MOON	$-4.22\,[-5.28, -3.18]$	$+2.24\,[+0.98, +3.52]$	$-6.45\,[-7.78, -5.12]$
MOON (soft)	$-5.93\,[-7.47, -4.49]$	$+1.81\,[+0.47, +3.14]$	$-7.74\,[-9.33, -6.15]$

Table 49: Paired matched-context contrast (random minus far), LAION ViT-B/32, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$+0.00\,[+0.00, +0.00]$	$-0.66\,[-1.52, +0.18]$	$+0.66\,[-0.18, +1.52]$
ZLaP	$+3.26\,[+1.89, +4.64]$	$+2.30\,[+1.16, +3.48]$	$+0.96\,[-0.78, +2.71]$
GDA-CLIP	$+7.90\,[+6.45, +9.36]$	$+7.06\,[+5.49, +8.59]$	$+0.83\,[-1.33, +3.00]$
TransCLIP	$+6.90\,[+5.55, +8.30]$	$-2.69\,[-4.25, -1.20]$	$+9.59\,[+7.69, +11.49]$
ADAPT	$+0.15\,[-0.06, +0.39]$	$-0.26\,[-1.65, +1.10]$	$+0.42\,[-0.98, +1.81]$
StatA	$-2.28\,[-3.24, -1.45]$	$-1.72\,[-3.06, -0.41]$	$-0.56\,[-1.94, +0.80]$
MOON	$-2.31\,[-3.43, -1.26]$	$-1.34\,[-2.54, -0.16]$	$-0.96\,[-2.47, +0.50]$
MOON (soft)	$-1.23\,[-2.36, -0.13]$	$-1.13\,[-2.36, +0.07]$	$-0.10\,[-1.61, +1.35]$

Table 50: Paired matched-context contrast (near minus far), LAION ViT-B/32, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.03\,[-0.09, +0.04]$	$+0.00\,[-0.27, +0.28]$	$-0.03\,[-0.30, +0.25]$
ZLaP	$+2.11\,[+1.19, +3.03]$	$+0.17\,[-0.36, +0.70]$	$+1.94\,[+0.92, +2.97]$
GDA-CLIP	$+7.36\,[+6.45, +8.26]$	$+3.36\,[+2.77, +3.93]$	$+4.00\,[+2.96, +5.08]$
TransCLIP	$+7.01\,[+6.19, +7.84]$	$-0.73\,[-1.37, -0.08]$	$+7.74\,[+6.81, +8.68]$
ADAPT	$-0.44\,[-0.62, -0.25]$	$+0.77\,[+0.24, +1.29]$	$-1.21\,[-1.71, -0.69]$
StatA	$-0.60\,[-1.14, -0.07]$	$+0.23\,[-0.29, +0.75]$	$-0.83\,[-1.39, -0.27]$
MOON	$-2.22\,[-2.74, -1.72]$	$-0.11\,[-0.54, +0.31]$	$-2.12\,[-2.69, -1.54]$
MOON (soft)	$-2.28\,[-2.85, -1.71]$	$-0.07\,[-0.51, +0.35]$	$-2.21\,[-2.81, -1.61]$

Table 51: Paired matched-context contrast (near minus random), LAION ViT-B/32, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$+0.01\,[-0.05, +0.06]$	$+0.37\,[+0.08, +0.66]$	$-0.37\,[-0.66, -0.07]$
ZLaP	$+1.92\,[+1.02, +2.84]$	$+0.74\,[+0.18, +1.29]$	$+1.18\,[+0.15, +2.24]$
GDA-CLIP	$+6.52\,[+5.63, +7.39]$	$+3.49\,[+2.94, +4.03]$	$+3.04\,[+1.96, +4.10]$
TransCLIP	$+6.26\,[+5.47, +7.04]$	$+0.93\,[+0.23, +1.64]$	$+5.33\,[+4.36, +6.29]$
ADAPT	$-0.34\,[-0.54, -0.14]$	$+2.07\,[+1.49, +2.64]$	$-2.41\,[-2.94, -1.86]$
StatA	$-2.10\,[-2.65, -1.54]$	$+0.42\,[-0.09, +0.91]$	$-2.51\,[-3.12, -1.88]$
MOON	$-2.95\,[-3.51, -2.40]$	$+0.76\,[+0.24, +1.28]$	$-3.71\,[-4.37, -3.07]$
MOON (soft)	$-3.25\,[-3.89, -2.61]$	$+0.54\,[+0.03, +1.05]$	$-3.79\,[-4.50, -3.07]$

Table 52: Paired matched-context contrast (random minus far), LAION ViT-B/32, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.03\,[-0.09, +0.03]$	$-0.37\,[-0.65, -0.10]$	$+0.34\,[+0.07, +0.61]$
ZLaP	$+0.19\,[-0.67, +1.03]$	$-0.58\,[-1.01, -0.14]$	$+0.76\,[-0.18, +1.69]$
GDA-CLIP	$+0.83\,[-0.03, +1.71]$	$-0.13\,[-0.66, +0.40]$	$+0.96\,[-0.07, +2.00]$
TransCLIP	$+0.75\,[-0.04, +1.55]$	$-1.66\,[-2.19, -1.14]$	$+2.41\,[+1.52, +3.30]$
ADAPT	$-0.10\,[-0.27, +0.07]$	$-1.30\,[-1.82, -0.77]$	$+1.20\,[+0.69, +1.72]$
StatA	$+1.50\,[+1.02, +1.97]$	$-0.19\,[-0.69, +0.32]$	$+1.68\,[+1.08, +2.26]$
MOON	$+0.72\,[+0.20, +1.25]$	$-0.87\,[-1.34, -0.40]$	$+1.59\,[+0.98, +2.21]$
MOON (soft)	$+0.97\,[+0.42, +1.52]$	$-0.61\,[-1.07, -0.17]$	$+1.58\,[+0.93, +2.23]$

Table 53: Paired matched-context contrast (near minus far), OpenAI ViT-B/16, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.01\,[-0.21, +0.20]$	$+0.72\,[-0.07, +1.50]$	$-0.73\,[-1.49, +0.04]$
ZLaP	$+8.81\,[+6.70, +10.82]$	$+2.88\,[+1.04, +4.80]$	$+5.93\,[+3.46, +8.37]$
GDA-CLIP	$+19.15\,[+17.20, +21.10]$	$+12.38\,[+10.60, +14.24]$	$+6.76\,[+4.09, +9.37]$
TransCLIP	$+16.15\,[+14.27, +18.06]$	$+1.07\,[-0.85, +2.99]$	$+15.08\,[+12.73, +17.35]$
ADAPT	$-1.22\,[-2.01, -0.46]$	$+5.50\,[+3.72, +7.34]$	$-6.72\,[-8.55, -5.03]$
StatA	$-7.04\,[-8.55, -5.60]$	$-0.37\,[-2.10, +1.34]$	$-6.68\,[-8.58, -4.81]$
MOON	$-8.48\,[-10.20, -6.86]$	$+2.10\,[+0.36, +3.86]$	$-10.57\,[-12.60, -8.54]$
MOON (soft)	$-9.59\,[-11.52, -7.76]$	$+1.11\,[-0.60, +2.88]$	$-10.70\,[-12.75, -8.69]$

Table 54: Paired matched-context contrast (near minus random), OpenAI ViT-B/16, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.10\,[-0.31, +0.12]$	$+1.98\,[+0.80, +3.21]$	$-2.08\,[-3.31, -0.92]$
ZLaP	$+4.97\,[+3.07, +6.88]$	$+1.11\,[-0.61, +2.87]$	$+3.86\,[+1.53, +6.18]$
GDA-CLIP	$+11.07\,[+9.15, +12.93]$	$+5.83\,[+4.30, +7.40]$	$+5.25\,[+2.94, +7.52]$
TransCLIP	$+9.92\,[+8.31, +11.53]$	$+3.19\,[+1.07, +5.30]$	$+6.73\,[+4.28, +9.22]$
ADAPT	$-0.93\,[-1.68, -0.22]$	$+5.72\,[+3.84, +7.67]$	$-6.65\,[-8.47, -4.94]$
StatA	$-5.37\,[-6.82, -3.99]$	$+2.54\,[+0.75, +4.35]$	$-7.91\,[-9.81, -6.07]$
MOON	$-5.87\,[-7.40, -4.46]$	$+4.93\,[+3.11, +6.76]$	$-10.80\,[-12.75, -8.93]$
MOON (soft)	$-6.69\,[-8.37, -5.04]$	$+4.62\,[+2.77, +6.52]$	$-11.31\,[-13.23, -9.41]$

Table 55: Paired matched-context contrast (random minus far), OpenAI ViT-B/16, $Q = 8$ fixed queries. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$+0.09\,[-0.03, +0.21]$	$-1.27\,[-2.50, -0.08]$	$+1.35\,[+0.16, +2.59]$
ZLaP	$+3.83\,[+2.13, +5.55]$	$+1.77\,[+0.39, +3.19]$	$+2.07\,[-0.11, +4.24]$
GDA-CLIP	$+8.07\,[+6.53, +9.64]$	$+6.55\,[+5.08, +8.06]$	$+1.52\,[-0.56, +3.65]$
TransCLIP	$+6.23\,[+4.76, +7.71]$	$-2.12\,[-4.23, -0.04]$	$+8.35\,[+6.07, +10.61]$
ADAPT	$-0.29\,[-0.75, +0.15]$	$-0.22\,[-2.04, +1.64]$	$-0.07\,[-1.95, +1.77]$
StatA	$-1.68\,[-2.51, -0.85]$	$-2.91\,[-4.79, -1.20]$	$+1.23\,[-0.57, +3.18]$
MOON	$-2.61\,[-3.63, -1.70]$	$-2.83\,[-4.66, -1.07]$	$+0.23\,[-1.63, +2.13]$
MOON (soft)	$-2.90\,[-3.93, -1.95]$	$-3.51\,[-5.27, -1.80]$	$+0.61\,[-1.26, +2.49]$

Table 56: Paired matched-context contrast (near minus far), OpenAI ViT-B/16, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.29\,[-0.40, -0.18]$	$-0.74\,[-1.07, -0.41]$	$+0.45\,[+0.12, +0.77]$
ZLaP	$+2.53\,[+1.54, +3.51]$	$+0.78\,[+0.26, +1.31]$	$+1.75\,[+0.76, +2.74]$
GDA-CLIP	$+7.19\,[+6.25, +8.11]$	$+4.49\,[+3.89, +5.07]$	$+2.70\,[+1.68, +3.71]$
TransCLIP	$+7.61\,[+6.67, +8.54]$	$+0.28\,[-0.43, +0.98]$	$+7.33\,[+6.34, +8.31]$
ADAPT	$-1.05\,[-1.37, -0.74]$	$+0.42\,[-0.23, +1.07]$	$-1.47\,[-2.08, -0.86]$
StatA	$-0.32\,[-0.98, +0.31]$	$-0.41\,[-0.98, +0.17]$	$+0.08\,[-0.51, +0.67]$
MOON	$-3.91\,[-4.53, -3.31]$	$-1.87\,[-2.49, -1.26]$	$-2.04\,[-2.72, -1.37]$
MOON (soft)	$-3.69\,[-4.35, -3.02]$	$-1.91\,[-2.56, -1.27]$	$-1.79\,[-2.48, -1.11]$

Table 57: Paired matched-context contrast (near minus random), OpenAI ViT-B/16, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.17\,[-0.28, -0.07]$	$+0.92\,[+0.56, +1.28]$	$-1.10\,[-1.45, -0.74]$
ZLaP	$+0.78\,[-0.28, +1.85]$	$+1.57\,[+1.03, +2.10]$	$-0.79\,[-1.93, +0.38]$
GDA-CLIP	$+5.26\,[+4.26, +6.26]$	$+3.99\,[+3.40, +4.59]$	$+1.27\,[+0.16, +2.37]$
TransCLIP	$+5.81\,[+4.93, +6.72]$	$+1.98\,[+1.19, +2.78]$	$+3.83\,[+2.77, +4.90]$
ADAPT	$-0.52\,[-0.85, -0.19]$	$+4.04\,[+3.32, +4.74]$	$-4.56\,[-5.25, -3.87]$
StatA	$-2.13\,[-2.81, -1.45]$	$+0.48\,[-0.11, +1.06]$	$-2.60\,[-3.29, -1.90]$
MOON	$-3.83\,[-4.42, -3.25]$	$+1.42\,[+0.78, +2.04]$	$-5.25\,[-6.04, -4.46]$
MOON (soft)	$-3.57\,[-4.24, -2.88]$	$+1.34\,[+0.64, +2.04]$	$-4.91\,[-5.76, -4.07]$

Table 58: Paired matched-context contrast (random minus far), OpenAI ViT-B/16, complete $N = 64$ episodes. Changes in full gain, rank gain, and obstruction relief are in percentage points; mean [95% CI].

Method	Full-gain contrast	Rank contrast	Relief contrast
Frozen CLIP	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$	$+0.00\,[+0.00, +0.00]$
Pseudo-linear	$-0.12\,[-0.21, -0.03]$	$-1.66\,[-2.01, -1.33]$	$+1.54\,[+1.20, +1.90]$
ZLaP	$+1.75\,[+0.84, +2.65]$	$-0.79\,[-1.22, -0.36]$	$+2.54\,[+1.54, +3.52]$
GDA-CLIP	$+1.93\,[+1.08, +2.77]$	$+0.50\,[-0.08, +1.07]$	$+1.43\,[+0.41, +2.43]$
TransCLIP	$+1.79\,[+1.04, +2.55]$	$-1.70\,[-2.41, -1.02]$	$+3.50\,[+2.52, +4.46]$
ADAPT	$-0.53\,[-0.84, -0.22]$	$-3.62\,[-4.33, -2.92]$	$+3.09\,[+2.36, +3.82]$
StatA	$+1.80\,[+1.18, +2.46]$	$-0.88\,[-1.36, -0.38]$	$+2.68\,[+2.00, +3.36]$
MOON	$-0.08\,[-0.77, +0.63]$	$-3.29\,[-3.92, -2.66]$	$+3.21\,[+2.39, +4.02]$
MOON (soft)	$-0.13\,[-0.89, +0.64]$	$-3.25\,[-3.90, -2.61]$	$+3.12\,[+2.24, +3.99]$

D.3 Changing context images while holding the class set fixed

This intervention retains the target queries and the identities of all eight represented classes. Only the 56 surrounding images change. High–low contrasts are paired by target class. The LAION replication contains Aircraft and Food and is therefore kept separate from the five-dataset OpenAI result.

E Additional support-sparsity results

The following tables give the complete fixed-N factorial results for $N \in \{32, 64, 96\}$, $K \in \{4, 8, 16, 32\}$, balanced and skewed class counts, and both OpenAI checkpoints. Each condition contains 1,000 shared episodes per dataset. The $N = 64$ results correspond to the main-text intervention; $N = 32$ and $N = 96$ test whether the same dependence on support size persists at smaller and larger batch sizes.

E.1 $N = 64$

E.2 $N = 32$

E.3 $N = 96$

Table 59: Fixed-class context-image intervention on five OpenAI ViT-B/32 development pools. The target queries and represented classes are identical across conditions. Full accuracies are percentages with pointwise 95% intervals; paired high–low contrasts are percentage points with 95% intervals.

Method	Full low	Full random	Full high	Full contrast	Rank contrast
Frozen CLIP	61.96 [59.63, 64.28]	61.96 [59.63, 64.28]	61.96 [59.63, 64.28]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
Pseudo-linear	62.04 [59.66, 64.40]	62.00 [59.65, 64.33]	62.38 [60.03, 64.70]	+0.34 [+0.06, +0.64]	+0.30 [−0.53, +1.15]
StatA	64.59 [61.97, 67.17]	62.57 [60.05, 65.03]	67.12 [64.72, 69.52]	+2.53 [+1.00, +4.24]	+3.41 [+1.68, +5.26]
MOON	62.81 [60.10, 65.48]	63.88 [61.24, 66.53]	69.24 [66.91, 71.61]	+6.43 [+5.23, +7.69]	+6.90 [+5.28, +8.60]
MOON (soft)	61.52 [58.63, 64.39]	64.51 [61.83, 67.15]	69.62 [67.36, 71.93]	+8.11 [+6.55, +9.84]	+8.40 [+6.81, +10.16]
ADAPT	61.80 [59.44, 64.16]	60.95 [58.51, 63.40]	60.36 [57.93, 62.79]	−1.44 [−2.36, −0.56]	−0.39 [−1.53, +0.74]

Table 60: Fixed-class context-image intervention on the Aircraft and Food LAION ViT-B/32 development pools. The target queries and represented classes are identical across conditions. Full accuracies are percentages with pointwise 95% intervals; paired high–low contrasts are percentage points with 95% intervals.

Method	Full low	Full random	Full high	Full contrast	Rank contrast
Frozen CLIP	51.35 [48.35, 54.41]	51.35 [48.35, 54.41]	51.35 [48.35, 54.41]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
Pseudo-linear	51.35 [48.28, 54.42]	51.29 [48.17, 54.35]	51.29 [48.18, 54.40]	−0.06 [−0.19, +0.00]	−0.06 [−1.13, +1.06]
StatA	54.50 [51.39, 57.62]	52.34 [49.15, 55.52]	54.14 [50.90, 57.38]	−0.36 [−1.93, +1.19]	−0.56 [−3.17, +1.94]
MOON	55.00 [51.51, 58.44]	54.64 [51.20, 58.07]	55.38 [52.02, 58.81]	+0.38 [−0.93, +1.81]	+1.44 [−0.37, +3.43]
MOON (soft)	54.81 [51.44, 58.13]	55.13 [51.82, 58.44]	57.12 [53.81, 60.49]	+2.31 [+0.63, +4.07]	+2.50 [+0.32, +4.76]
ADAPT	51.60 [48.54, 54.60]	51.04 [48.04, 54.04]	50.29 [47.24, 53.41]	−1.30 [−2.24, −0.37]	−0.49 [−2.05, +1.20]

Table 61: Fixed-$N = 64$ sparsity, OpenAI ViT-B/32, balanced class counts; five-dataset macro. Mean [95% CI]. Accuracies are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	62.46 [62.11, 62.81]	92.84 [92.62, 93.07]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.46 [62.11, 62.81]	31.17 [30.86, 31.46]	90.50 [90.11, 90.87]
4	MOON	72.84 [72.39, 73.29]	87.74 [87.41, 88.06]	−5.10 [−5.32, −4.89]	+15.49 [+15.19, +15.79]	71.89 [71.45, 72.34]	65.11 [64.60, 65.63]	78.83 [78.35, 79.31]
4	StatA	69.62 [69.21, 70.04]	90.85 [90.58, 91.13]	−1.99 [−2.13, −1.86]	+9.16 [+8.91, +9.42]	67.97 [67.56, 68.37]	45.97 [45.55, 46.40]	81.87 [81.41, 82.33]
4	ADAPT	61.03 [60.68, 61.38]	87.27 [86.92, 87.61]	−5.57 [−5.86, −5.28]	+4.15 [+3.85, +4.44]	62.46 [62.11, 62.81]	31.17 [30.86, 31.47]	90.46 [90.09, 90.84]
8	Frozen CLIP	62.22 [61.96, 62.48]	87.71 [87.51, 87.91]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.22 [61.96, 62.48]	41.06 [40.81, 41.30]	87.83 [87.55, 88.11]
8	MOON	70.36 [70.04, 70.68]	81.05 [80.78, 81.31]	−6.66 [−6.83, −6.48]	+14.80 [+14.58, +15.02]	69.90 [69.58, 70.21]	68.79 [68.40, 69.17]	76.18 [75.85, 76.52]
8	StatA	67.08 [66.78, 67.38]	84.30 [84.06, 84.53]	−3.41 [−3.54, −3.28]	+8.27 [+8.08, +8.47]	66.56 [66.26, 66.85]	52.84 [52.52, 53.16]	79.79 [79.47, 80.12]
8	ADAPT	62.06 [61.79, 62.32]	81.84 [81.58, 82.10]	−5.87 [−6.06, −5.67]	+5.71 [+5.51, +5.90]	62.22 [61.96, 62.48]	41.06 [40.82, 41.30]	87.82 [87.53, 88.10]
16	Frozen CLIP	62.13 [61.95, 62.32]	81.67 [81.50, 81.84]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.13 [61.95, 62.32]	53.85 [53.67, 54.04]	84.39 [84.17, 84.59]
16	MOON	65.79 [65.57, 66.01]	73.50 [73.30, 73.70]	−8.17 [−8.31, −8.02]	+11.82 [+11.66, +11.98]	66.11 [65.89, 66.33]	71.86 [71.58, 72.14]	72.96 [72.74, 73.19]
16	StatA	63.32 [63.11, 63.53]	77.00 [76.81, 77.19]	−4.67 [−4.79, −4.55]	+5.85 [+5.71, +5.99]	63.95 [63.75, 64.16]	59.89 [59.66, 60.11]	77.32 [77.09, 77.55]
16	ADAPT	62.14 [61.95, 62.33]	75.41 [75.21, 75.61]	−6.26 [−6.41, −6.11]	+6.26 [+6.12, +6.41]	62.13 [61.95, 62.33]	53.85 [53.67, 54.04]	84.38 [84.17, 84.59]
32	Frozen CLIP	62.10 [61.95, 62.25]	74.78 [74.64, 74.93]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.10 [61.95, 62.25]	68.76 [68.61, 68.91]	79.53 [79.37, 79.68]
32	MOON	59.95 [59.79, 60.11]	65.23 [65.08, 65.39]	−9.55 [−9.68, −9.43]	+7.40 [+7.29, +7.52]	60.82 [60.65, 60.98]	77.11 [76.90, 77.32]	67.64 [67.48, 67.81]
32	StatA	60.23 [60.07, 60.39]	69.24 [69.09, 69.39]	−5.55 [−5.65, −5.45]	+3.68 [+3.58, +3.79]	61.36 [61.20, 61.52]	70.43 [70.26, 70.60]	73.57 [73.40, 73.75]
32	ADAPT	62.12 [61.96, 62.27]	68.54 [68.38, 68.69]	−6.25 [−6.36, −6.14]	+6.27 [+6.16, +6.37]	62.10 [61.95, 62.25]	68.76 [68.61, 68.91]	79.52 [79.37, 79.68]

Table 62: Fixed-$N = 64$ sparsity, OpenAI ViT-B/32, skewed class counts; five-dataset macro. Mean [95% CI]. Accuracies are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	62.42 [62.00, 62.85]	92.73 [92.46, 93.01]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.42 [62.00, 62.85]	32.42 [32.10, 32.74]	88.76 [88.37, 89.16]
4	MOON	72.31 [71.78, 72.84]	88.68 [88.30, 89.06]	−4.06 [−4.31, −3.82]	+13.94 [+13.63, +14.26]	71.20 [70.68, 71.72]	64.18 [63.66, 64.71]	78.19 [77.72, 78.66]
4	StatA	69.43 [68.93, 69.93]	91.20 [90.87, 91.53]	−1.53 [−1.69, −1.37]	+8.53 [+8.25, +8.81]	67.58 [67.11, 68.06]	46.15 [45.71, 46.59]	80.88 [80.41, 81.34]
4	ADAPT	58.66 [58.27, 59.07]	87.35 [86.94, 87.76]	−5.38 [−5.71, −5.05]	+1.62 [+1.26, +1.99]	62.42 [62.00, 62.85]	32.42 [32.10, 32.74]	88.74 [88.35, 89.13]
8	Frozen CLIP	62.40 [62.09, 62.71]	87.77 [87.54, 88.01]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.40 [62.09, 62.71]	41.83 [41.56, 42.09]	85.47 [85.17, 85.77]
8	MOON	70.48 [70.10, 70.87]	82.38 [82.06, 82.69]	−5.39 [−5.59, −5.20]	+13.48 [+13.24, +13.72]	69.73 [69.35, 70.11]	67.55 [67.16, 67.94]	74.11 [73.76, 74.45]
8	StatA	67.81 [67.45, 68.16]	85.36 [85.08, 85.64]	−2.41 [−2.56, −2.27]	+7.82 [+7.62, +8.03]	66.70 [66.35, 67.04]	52.58 [52.25, 52.91]	78.03 [77.69, 78.37]
8	ADAPT	61.46 [61.16, 61.76]	82.43 [82.11, 82.73]	−5.35 [−5.58, −5.13]	+4.41 [+4.18, +4.64]	62.40 [62.09, 62.71]	41.83 [41.56, 42.09]	85.46 [85.15, 85.76]
16	Frozen CLIP	62.11 [61.89, 62.33]	81.60 [81.41, 81.80]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.11 [61.89, 62.33]	53.98 [53.77, 54.18]	81.57 [81.34, 81.79]
16	MOON	66.37 [66.11, 66.63]	74.82 [74.58, 75.05]	−6.79 [−6.94, −6.63]	+11.04 [+10.87, +11.21]	66.16 [65.90, 66.42]	71.58 [71.30, 71.87]	69.30 [69.06, 69.54]
16	StatA	64.45 [64.20, 64.70]	78.15 [77.94, 78.37]	−3.45 [−3.57, −3.32]	+5.78 [+5.63, +5.94]	64.37 [64.12, 64.62]	60.27 [60.03, 60.51]	74.59 [74.35, 74.84]
16	ADAPT	62.04 [61.81, 62.26]	76.02 [75.80, 76.25]	−5.58 [−5.74, −5.42]	+5.50 [+5.34, +5.66]	62.11 [61.89, 62.34]	53.98 [53.77, 54.19]	81.57 [81.34, 81.79]
32	Frozen CLIP	62.13 [61.97, 62.30]	74.79 [74.63, 74.95]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.13 [61.97, 62.30]	68.53 [68.38, 68.68]	76.67 [76.50, 76.84]
32	MOON	60.86 [60.67, 61.05]	66.29 [66.12, 66.46]	−8.50 [−8.62, −8.38]	+7.22 [+7.11, +7.34]	61.06 [60.87, 61.24]	77.56 [77.34, 77.78]	63.52 [63.34, 63.70]
32	StatA	61.27 [61.09, 61.45]	70.12 [69.95, 70.29]	−4.67 [−4.78, −4.56]	+3.80 [+3.70, +3.92]	61.74 [61.56, 61.92]	70.83 [70.65, 71.01]	70.35 [70.17, 70.53]
32	ADAPT	62.14 [61.98, 62.31]	69.16 [68.99, 69.32]	−5.63 [−5.75, −5.52]	+5.64 [+5.53, +5.76]	62.14 [61.97, 62.30]	68.53 [68.38, 68.68]	76.67 [76.50, 76.84]

Table 63: Fixed-$N = 64$ sparsity, OpenAI ViT-B/16, balanced class counts; five-dataset macro. Mean [95% CI]. Accuracies are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	65.58 [65.23, 65.94]	94.20 [93.99, 94.42]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.58 [65.23, 65.94]	35.77 [35.43, 36.10]	90.03 [89.64, 90.40]
4	MOON	74.78 [74.33, 75.22]	89.17 [88.85, 89.48]	−5.03 [−5.25, −4.82]	+14.23 [+13.93, +14.54]	73.92 [73.48, 74.36]	68.44 [67.94, 68.95]	79.84 [79.37, 80.29]
4	StatA	72.03 [71.61, 72.46]	92.08 [91.82, 92.36]	−2.12 [−2.26, −1.97]	+8.56 [+8.32, +8.81]	70.79 [70.38, 71.21]	52.31 [51.86, 52.78]	81.54 [81.08, 82.00]
4	ADAPT	64.15 [63.79, 64.49]	87.69 [87.32, 88.04]	−6.51 [−6.83, −6.21]	+5.07 [+4.76, +5.40]	65.59 [65.23, 65.94]	35.77 [35.44, 36.11]	90.02 [89.63, 90.39]
8	Frozen CLIP	65.44 [65.17, 65.70]	89.76 [89.56, 89.96]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.44 [65.17, 65.70]	45.44 [45.18, 45.71]	87.74 [87.45, 88.02]
8	MOON	72.55 [72.22, 72.87]	83.16 [82.89, 83.42]	−6.60 [−6.79, −6.43]	+13.71 [+13.48, +13.95]	72.13 [71.81, 72.46]	72.03 [71.66, 72.41]	77.11 [76.78, 77.44]
8	StatA	70.06 [69.74, 70.37]	86.27 [86.03, 86.51]	−3.49 [−3.63, −3.36]	+8.11 [+7.91, +8.31]	69.66 [69.35, 69.97]	58.77 [58.43, 59.11]	79.68 [79.35, 80.02]
8	ADAPT	65.20 [64.93, 65.46]	83.29 [83.02, 83.56]	−6.47 [−6.68, −6.25]	+6.23 [+6.01, +6.45]	65.44 [65.17, 65.70]	45.44 [45.18, 45.72]	87.73 [87.44, 88.02]
16	Frozen CLIP	65.52 [65.32, 65.72]	84.26 [84.09, 84.44]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.52 [65.32, 65.72]	57.80 [57.59, 58.00]	84.90 [84.69, 85.11]
16	MOON	68.53 [68.30, 68.76]	76.47 [76.27, 76.68]	−7.79 [−7.93, −7.65]	+10.80 [+10.63, +10.97]	68.78 [68.55, 69.01]	74.80 [74.52, 75.07]	74.07 [73.84, 74.30]
16	StatA	66.86 [66.63, 67.08]	79.51 [79.32, 79.70]	−4.75 [−4.87, −4.64]	+6.09 [+5.94, +6.23]	67.31 [67.09, 67.53]	64.77 [64.53, 65.01]	77.64 [77.40, 77.88]
16	ADAPT	65.53 [65.33, 65.72]	78.14 [77.94, 78.34]	−6.12 [−6.28, −5.97]	+6.13 [+5.98, +6.28]	65.52 [65.32, 65.72]	57.80 [57.59, 58.00]	84.89 [84.68, 85.10]
32	Frozen CLIP	65.59 [65.44, 65.74]	77.88 [77.74, 78.02]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.59 [65.44, 65.74]	71.62 [71.47, 71.77]	80.56 [80.40, 80.72]
32	MOON	62.95 [62.78, 63.11]	68.60 [68.44, 68.76]	−9.28 [−9.40, −9.16]	+6.64 [+6.52, +6.75]	63.79 [63.62, 63.95]	79.20 [78.99, 79.40]	69.29 [69.12, 69.45]
32	StatA	63.74 [63.58, 63.90]	72.34 [72.19, 72.49]	−5.55 [−5.65, −5.44]	+3.70 [+3.59, +3.81]	64.74 [64.57, 64.90]	73.51 [73.34, 73.68]	74.49 [74.32, 74.66]
32	ADAPT	65.61 [65.46, 65.76]	71.85 [71.70, 72.00]	−6.04 [−6.14, −5.93]	+6.06 [+5.95, +6.17]	65.59 [65.44, 65.74]	71.63 [71.47, 71.77]	80.56 [80.40, 80.72]

Table 64: Fixed-$N = 64$ sparsity, OpenAI ViT-B/16, skewed class counts; five-dataset macro. Mean [95% CI]. Accuracies are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	65.66 [65.23, 66.09]	94.24 [93.99, 94.48]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.66 [65.23, 66.09]	37.40 [37.04, 37.75]	88.55 [88.14, 88.95]
4	MOON	74.17 [73.66, 74.69]	90.16 [89.79, 90.51]	−4.08 [−4.33, −3.84]	+12.59 [+12.28, +12.91]	73.19 [72.68, 73.70]	67.45 [66.94, 67.95]	79.03 [78.56, 79.49]
4	StatA	71.92 [71.42, 72.43]	92.48 [92.16, 92.79]	−1.76 [−1.92, −1.59]	+8.02 [+7.75, +8.30]	70.50 [70.02, 70.99]	52.62 [52.16, 53.10]	80.84 [80.37, 81.30]
4	ADAPT	62.10 [61.68, 62.50]	87.97 [87.54, 88.39]	−6.27 [−6.64, −5.91]	+2.71 [+2.32, +3.10]	65.66 [65.23, 66.09]	37.40 [37.04, 37.76]	88.54 [88.14, 88.94]
8	Frozen CLIP	65.68 [65.36, 65.99]	89.84 [89.61, 90.07]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.68 [65.36, 65.99]	46.44 [46.16, 46.73]	85.58 [85.27, 85.88]
8	MOON	72.77 [72.39, 73.14]	84.35 [84.03, 84.66]	−5.49 [−5.68, −5.29]	+12.58 [+12.34, +12.83]	72.06 [71.68, 72.42]	70.79 [70.41, 71.17]	75.29 [74.95, 75.62]
8	StatA	70.51 [70.15, 70.88]	87.09 [86.81, 87.37]	−2.75 [−2.91, −2.60]	+7.59 [+7.37, +7.80]	69.67 [69.31, 70.03]	58.08 [57.73, 58.42]	78.17 [77.83, 78.51]
8	ADAPT	64.67 [64.36, 64.98]	83.82 [83.50, 84.13]	−6.02 [−6.27, −5.77]	+5.01 [+4.76, +5.27]	65.68 [65.36, 65.99]	46.45 [46.16, 46.73]	85.58 [85.27, 85.87]
16	Frozen CLIP	65.44 [65.21, 65.67]	84.23 [84.03, 84.43]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.44 [65.21, 65.67]	57.97 [57.75, 58.18]	82.29 [82.05, 82.51]
16	MOON	69.06 [68.79, 69.33]	77.59 [77.34, 77.83]	−6.65 [−6.80, −6.49]	+10.27 [+10.08, +10.45]	68.78 [68.51, 69.05]	74.33 [74.05, 74.60]	70.84 [70.60, 71.09]
16	StatA	67.81 [67.54, 68.08]	80.55 [80.32, 80.77]	−3.68 [−3.81, −3.55]	+6.05 [+5.89, +6.22]	67.71 [67.45, 67.97]	65.09 [64.84, 65.34]	75.47 [75.23, 75.72]
16	ADAPT	65.34 [65.11, 65.57]	78.43 [78.19, 78.66]	−5.80 [−5.96, −5.63]	+5.70 [+5.54, +5.87]	65.44 [65.21, 65.67]	57.97 [57.75, 58.18]	82.28 [82.05, 82.51]
32	Frozen CLIP	65.55 [65.39, 65.72]	77.85 [77.69, 78.00]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.55 [65.39, 65.72]	71.43 [71.27, 71.59]	77.93 [77.77, 78.10]
32	MOON	63.88 [63.69, 64.07]	69.61 [69.43, 69.79]	−8.24 [−8.36, −8.11]	+6.57 [+6.44, +6.69]	64.04 [63.86, 64.23]	79.69 [79.48, 79.90]	65.66 [65.48, 65.84]
32	StatA	64.80 [64.62, 64.98]	73.18 [73.01, 73.35]	−4.67 [−4.78, −4.56]	+3.92 [+3.81, +4.03]	65.08 [64.90, 65.26]	73.99 [73.81, 74.16]	71.62 [71.44, 71.80]
32	ADAPT	65.56 [65.39, 65.72]	72.21 [72.04, 72.39]	−5.64 [−5.75, −5.52]	+5.64 [+5.53, +5.76]	65.55 [65.39, 65.72]	71.43 [71.27, 71.59]	77.93 [77.76, 78.10]

Table 65: Fixed-$N = 32$ support-sparsity intervention, OpenAI ViT-B/32, balanced class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	62.01 [61.64, 62.39]	93.01 [92.78, 93.24]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.01 [61.64, 62.39]	39.86 [39.50, 40.23]	86.17 [85.73, 86.59]
4	MOON	71.63 [71.16, 72.09]	87.98 [87.65, 88.29]	−5.04 [−5.26, −4.82]	+14.65 [+14.32, +14.98]	71.22 [70.75, 71.69]	67.18 [66.66, 67.69]	76.47 [76.00, 76.94]
4	StatA	67.67 [67.24, 68.09]	91.03 [90.75, 91.31]	−1.98 [−2.12, −1.85]	+7.64 [+7.37, +7.90]	67.19 [66.77, 67.61]	51.94 [51.48, 52.39]	79.55 [79.08, 80.00]
4	ADAPT	61.82 [61.45, 62.19]	83.85 [83.46, 84.22]	−9.17 [−9.50, −8.84]	+8.97 [+8.64, +9.31]	62.01 [61.64, 62.39]	39.86 [39.50, 40.23]	86.17 [85.73, 86.58]
8	Frozen CLIP	61.97 [61.68, 62.25]	87.70 [87.48, 87.91]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.97 [61.68, 62.25]	50.11 [49.82, 50.40]	82.39 [82.07, 82.71]
8	MOON	67.73 [67.40, 68.07]	80.66 [80.38, 80.95]	−7.04 [−7.23, −6.84]	+12.80 [+12.55, +13.05]	67.89 [67.56, 68.22]	68.04 [67.66, 68.43]	73.86 [73.53, 74.21]
8	StatA	64.30 [63.99, 64.62]	84.23 [83.98, 84.48]	−3.47 [−3.61, −3.33]	+5.80 [+5.60, +6.01]	64.80 [64.48, 65.12]	56.20 [55.85, 56.54]	77.43 [77.09, 77.77]
8	ADAPT	61.95 [61.66, 62.24]	77.61 [77.31, 77.91]	−10.09 [−10.34, −9.84]	+10.08 [+9.83, +10.33]	61.97 [61.68, 62.25]	50.11 [49.82, 50.40]	82.39 [82.06, 82.71]
16	Frozen CLIP	62.00 [61.77, 62.22]	81.52 [81.33, 81.72]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.00 [61.77, 62.22]	61.62 [61.40, 61.85]	76.75 [76.51, 76.98]
16	MOON	61.96 [61.72, 62.20]	72.53 [72.30, 72.76]	−8.99 [−9.17, −8.82]	+8.95 [+8.76, +9.15]	62.63 [62.38, 62.87]	69.76 [69.47, 70.05]	69.14 [68.89, 69.38]
16	StatA	60.95 [60.71, 61.17]	76.83 [76.63, 77.04]	−4.69 [−4.83, −4.55]	+3.63 [+3.46, +3.80]	62.02 [61.78, 62.25]	62.61 [62.37, 62.85]	73.30 [73.05, 73.55]
16	ADAPT	62.00 [61.77, 62.22]	70.47 [70.24, 70.70]	−11.05 [−11.25, −10.86]	+11.05 [+10.86, +11.25]	62.00 [61.77, 62.22]	61.62 [61.40, 61.85]	76.75 [76.51, 76.98]
32	Frozen CLIP	62.08 [61.89, 62.28]	74.63 [74.45, 74.81]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.08 [61.89, 62.28]	73.56 [73.37, 73.74]	67.14 [66.96, 67.32]
32	MOON	58.48 [58.29, 58.67]	65.03 [64.84, 65.21]	−9.60 [−9.76, −9.46]	+6.00 [+5.86, +6.14]	58.62 [58.43, 58.81]	76.85 [76.62, 77.08]	60.42 [60.23, 60.61]
32	StatA	60.12 [59.92, 60.31]	69.38 [69.20, 69.57]	−5.25 [−5.38, −5.12]	+3.28 [+3.14, +3.42]	59.89 [59.69, 60.08]	73.28 [73.09, 73.47]	64.36 [64.17, 64.55]
32	ADAPT	62.08 [61.89, 62.28]	64.01 [63.82, 64.20]	−10.62 [−10.78, −10.47]	+10.62 [+10.47, +10.78]	62.08 [61.89, 62.28]	73.56 [73.37, 73.74]	67.14 [66.96, 67.32]

Table 66: Fixed-$N = 32$ support-sparsity intervention, OpenAI ViT-B/32, skewed class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	61.96 [61.52, 62.41]	92.93 [92.64, 93.20]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.96 [61.52, 62.41]	40.76 [40.37, 41.15]	83.52 [83.06, 83.97]
4	MOON	71.31 [70.76, 71.86]	89.20 [88.82, 89.57]	−3.73 [−3.97, −3.49]	+13.07 [+12.73, +13.41]	70.50 [69.96, 71.03]	65.27 [64.75, 65.79]	74.97 [74.49, 75.45]
4	StatA	68.01 [67.49, 68.51]	91.72 [91.40, 92.05]	−1.20 [−1.35, −1.05]	+7.25 [+6.95, +7.54]	66.96 [66.46, 67.46]	51.72 [51.26, 52.19]	77.50 [77.01, 77.97]
4	ADAPT	61.00 [60.56, 61.43]	84.79 [84.35, 85.24]	−8.13 [−8.52, −7.74]	+7.17 [+6.76, +7.57]	61.96 [61.52, 62.41]	40.76 [40.37, 41.15]	83.51 [83.05, 83.95]
8	Frozen CLIP	61.94 [61.61, 62.27]	87.66 [87.42, 87.91]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	61.94 [61.61, 62.27]	50.23 [49.93, 50.54]	79.19 [78.86, 79.53]
8	MOON	68.31 [67.91, 68.70]	81.94 [81.63, 82.25]	−5.73 [−5.94, −5.51]	+12.09 [+11.82, +12.36]	67.94 [67.55, 68.34]	67.54 [67.14, 67.95]	70.70 [70.34, 71.06]
8	StatA	65.31 [64.94, 65.67]	85.22 [84.95, 85.50]	−2.44 [−2.60, −2.29]	+5.81 [+5.58, +6.04]	65.16 [64.79, 65.52]	56.57 [56.21, 56.92]	74.30 [73.94, 74.65]
8	ADAPT	61.86 [61.53, 62.18]	78.37 [78.04, 78.70]	−9.29 [−9.57, −9.01]	+9.20 [+8.92, +9.48]	61.94 [61.62, 62.27]	50.23 [49.93, 50.54]	79.19 [78.85, 79.52]
16	Frozen CLIP	62.03 [61.78, 62.28]	81.63 [81.42, 81.84]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.03 [61.78, 62.28]	61.37 [61.13, 61.61]	73.64 [73.39, 73.89]
16	MOON	63.22 [62.94, 63.50]	73.95 [73.69, 74.20]	−7.68 [−7.86, −7.50]	+8.87 [+8.67, +9.07]	63.31 [63.03, 63.59]	70.50 [70.20, 70.80]	65.39 [65.11, 65.66]
16	StatA	62.07 [61.80, 62.34]	77.74 [77.50, 77.98]	−3.89 [−4.03, −3.74]	+3.92 [+3.74, +4.10]	62.53 [62.26, 62.79]	63.27 [63.02, 63.52]	69.83 [69.57, 70.09]
16	ADAPT	62.04 [61.79, 62.29]	71.42 [71.17, 71.67]	−10.21 [−10.42, −10.00]	+10.21 [+10.00, +10.42]	62.03 [61.78, 62.28]	61.37 [61.13, 61.61]	73.64 [73.39, 73.89]
32	Frozen CLIP	62.08 [61.89, 62.28]	74.63 [74.45, 74.81]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.08 [61.89, 62.28]	73.56 [73.37, 73.74]	67.14 [66.96, 67.32]
32	MOON	58.48 [58.29, 58.67]	65.03 [64.84, 65.21]	−9.60 [−9.76, −9.46]	+6.00 [+5.86, +6.14]	58.62 [58.43, 58.81]	76.85 [76.62, 77.08]	60.42 [60.23, 60.61]
32	StatA	60.12 [59.92, 60.31]	69.38 [69.20, 69.57]	−5.25 [−5.38, −5.12]	+3.28 [+3.14, +3.42]	59.89 [59.69, 60.08]	73.28 [73.09, 73.47]	64.36 [64.17, 64.55]
32	ADAPT	62.08 [61.89, 62.28]	64.01 [63.82, 64.20]	−10.62 [−10.78, −10.47]	+10.62 [+10.47, +10.78]	62.08 [61.89, 62.28]	73.56 [73.37, 73.74]	67.14 [66.96, 67.32]

Table 67: Fixed-$N = 32$ support-sparsity intervention, OpenAI ViT-B/16, balanced class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	65.61 [65.23, 65.99]	94.30 [94.07, 94.51]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.61 [65.23, 65.99]	44.62 [44.22, 45.02]	86.98 [86.57, 87.40]
4	MOON	74.17 [73.70, 74.63]	89.34 [89.02, 89.66]	−4.95 [−5.17, −4.74]	+13.51 [+13.19, +13.84]	73.77 [73.31, 74.23]	70.96 [70.45, 71.48]	78.00 [77.52, 78.47]
4	StatA	71.04 [70.61, 71.49]	92.19 [91.91, 92.46]	−2.11 [−2.25, −1.96]	+7.54 [+7.29, +7.80]	70.72 [70.29, 71.16]	58.21 [57.73, 58.69]	80.17 [79.70, 80.63]
4	ADAPT	65.34 [64.97, 65.72]	85.15 [84.76, 85.54]	−9.15 [−9.50, −8.80]	+8.89 [+8.54, +9.24]	65.61 [65.23, 65.99]	44.62 [44.22, 45.02]	86.98 [86.57, 87.40]
8	Frozen CLIP	65.49 [65.21, 65.77]	89.73 [89.52, 89.95]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.49 [65.21, 65.77]	54.57 [54.27, 54.87]	83.64 [83.33, 83.95]
8	MOON	70.46 [70.12, 70.79]	83.22 [82.95, 83.49]	−6.51 [−6.70, −6.32]	+11.48 [+11.23, +11.73]	70.57 [70.23, 70.90]	71.56 [71.17, 71.94]	75.22 [74.88, 75.55]
8	StatA	67.99 [67.68, 68.30]	86.38 [86.14, 86.63]	−3.35 [−3.48, −3.21]	+5.84 [+5.64, +6.04]	68.34 [68.03, 68.65]	61.62 [61.27, 61.96]	78.16 [77.82, 78.49]
8	ADAPT	65.49 [65.21, 65.77]	79.94 [79.65, 80.23]	−9.78 [−10.04, −9.54]	+9.78 [+9.54, +10.03]	65.49 [65.21, 65.77]	54.57 [54.27, 54.88]	83.64 [83.33, 83.95]
16	Frozen CLIP	65.50 [65.28, 65.72]	84.23 [84.04, 84.41]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.50 [65.28, 65.72]	65.15 [64.93, 65.38]	78.49 [78.25, 78.72]
16	MOON	65.09 [64.85, 65.34]	76.04 [75.81, 76.27]	−8.19 [−8.36, −8.02]	+7.79 [+7.61, +7.97]	65.65 [65.41, 65.89]	72.72 [72.44, 73.01]	71.03 [70.78, 71.27]
16	StatA	64.72 [64.49, 64.95]	79.66 [79.46, 79.86]	−4.57 [−4.71, −4.44]	+3.79 [+3.63, +3.95]	65.60 [65.37, 65.84]	66.67 [66.44, 66.92]	74.91 [74.66, 75.15]
16	ADAPT	65.50 [65.28, 65.72]	73.63 [73.41, 73.86]	−10.60 [−10.79, −10.41]	+10.60 [+10.42, +10.80]	65.50 [65.28, 65.72]	65.15 [64.93, 65.38]	78.49 [78.25, 78.72]
32	Frozen CLIP	65.54 [65.35, 65.72]	77.70 [77.53, 77.87]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.54 [65.35, 65.72]	76.03 [75.85, 76.21]	69.62 [69.45, 69.80]
32	MOON	61.59 [61.40, 61.78]	68.50 [68.31, 68.68]	−9.21 [−9.36, −9.05]	+5.26 [+5.12, +5.40]	61.76 [61.57, 61.95]	79.09 [78.87, 79.31]	63.17 [62.98, 63.36]
32	StatA	63.80 [63.60, 63.99]	72.74 [72.57, 72.92]	−4.96 [−5.09, −4.83]	+3.23 [+3.09, +3.36]	63.57 [63.38, 63.76]	76.08 [75.89, 76.27]	66.98 [66.80, 67.17]
32	ADAPT	65.54 [65.35, 65.72]	67.24 [67.06, 67.43]	−10.46 [−10.61, −10.31]	+10.46 [+10.31, +10.61]	65.54 [65.35, 65.72]	76.03 [75.85, 76.21]	69.62 [69.45, 69.80]

Table 68: Fixed-$N = 32$ support-sparsity intervention, OpenAI ViT-B/16, skewed class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	65.49 [65.05, 65.95]	94.32 [94.06, 94.59]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.49 [65.05, 65.95]	45.70 [45.27, 46.13]	84.51 [84.07, 84.95]
4	MOON	73.99 [73.44, 74.55]	90.47 [90.10, 90.84]	−3.85 [−4.09, −3.61]	+12.35 [+12.00, +12.70]	73.24 [72.70, 73.79]	69.04 [68.51, 69.55]	76.65 [76.17, 77.12]
4	StatA	71.20 [70.67, 71.73]	92.78 [92.46, 93.09]	−1.55 [−1.70, −1.39]	+7.25 [+6.96, +7.55]	70.47 [69.96, 70.97]	57.80 [57.31, 58.28]	78.55 [78.07, 79.03]
4	ADAPT	64.49 [64.04, 64.93]	85.95 [85.50, 86.42]	−8.37 [−8.77, −7.97]	+7.36 [+6.95, +7.77]	65.49 [65.05, 65.95]	45.70 [45.28, 46.13]	84.50 [84.07, 84.95]
8	Frozen CLIP	65.40 [65.07, 65.72]	89.73 [89.49, 89.97]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.40 [65.07, 65.72]	54.66 [54.35, 54.98]	80.82 [80.49, 81.14]
8	MOON	70.95 [70.55, 71.35]	84.39 [84.09, 84.70]	−5.34 [−5.54, −5.13]	+10.88 [+10.62, +11.15]	70.63 [70.24, 71.02]	71.06 [70.67, 71.44]	72.46 [72.11, 72.81]
8	StatA	68.64 [68.26, 69.01]	87.07 [86.79, 87.35]	−2.66 [−2.81, −2.51]	+5.90 [+5.68, +6.13]	68.61 [68.24, 68.98]	61.97 [61.62, 62.31]	75.55 [75.20, 75.88]
8	ADAPT	65.30 [64.97, 65.62]	80.47 [80.13, 80.81]	−9.27 [−9.55, −8.99]	+9.16 [+8.89, +9.45]	65.40 [65.07, 65.72]	54.66 [54.35, 54.98]	80.82 [80.49, 81.14]
16	Frozen CLIP	65.68 [65.44, 65.93]	84.25 [84.05, 84.46]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.68 [65.44, 65.93]	65.10 [64.86, 65.34]	75.73 [75.48, 75.97]
16	MOON	66.41 [66.13, 66.68]	77.04 [76.79, 77.29]	−7.22 [−7.39, −7.05]	+7.94 [+7.75, +8.14]	66.48 [66.21, 66.75]	73.64 [73.34, 73.92]	67.85 [67.59, 68.11]
16	StatA	65.95 [65.69, 66.22]	80.33 [80.11, 80.56]	−3.92 [−4.06, −3.78]	+4.20 [+4.03, +4.37]	66.28 [66.02, 66.54]	67.56 [67.30, 67.82]	71.87 [71.62, 72.12]
16	ADAPT	65.68 [65.44, 65.93]	74.38 [74.13, 74.62]	−9.88 [−10.07, −9.68]	+9.88 [+9.68, +10.08]	65.68 [65.44, 65.93]	65.10 [64.86, 65.34]	75.73 [75.48, 75.97]
32	Frozen CLIP	65.54 [65.35, 65.72]	77.70 [77.53, 77.87]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.54 [65.35, 65.72]	76.03 [75.85, 76.21]	69.62 [69.45, 69.80]
32	MOON	61.59 [61.40, 61.78]	68.50 [68.31, 68.68]	−9.21 [−9.36, −9.05]	+5.26 [+5.12, +5.40]	61.76 [61.57, 61.95]	79.09 [78.87, 79.31]	63.17 [62.98, 63.36]
32	StatA	63.80 [63.60, 63.99]	72.74 [72.57, 72.92]	−4.96 [−5.09, −4.83]	+3.23 [+3.09, +3.36]	63.57 [63.38, 63.76]	76.08 [75.89, 76.27]	66.98 [66.80, 67.17]
32	ADAPT	65.54 [65.35, 65.72]	67.24 [67.06, 67.43]	−10.46 [−10.61, −10.31]	+10.46 [+10.31, +10.61]	65.54 [65.35, 65.72]	76.03 [75.85, 76.21]	69.62 [69.45, 69.80]

Table 69: Fixed-$N = 96$ support-sparsity intervention, OpenAI ViT-B/32, balanced class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	62.48 [62.13, 62.83]	92.98 [92.76, 93.21]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.48 [62.13, 62.83]	26.71 [26.45, 26.97]	91.81 [91.45, 92.15]
4	MOON	72.61 [72.16, 73.06]	87.87 [87.56, 88.19]	−5.11 [−5.32, −4.90]	+15.25 [+14.97, +15.53]	71.52 [71.08, 71.96]	62.09 [61.56, 62.61]	80.02 [79.55, 80.49]
4	StatA	69.82 [69.41, 70.23]	90.96 [90.69, 91.23]	−2.02 [−2.16, −1.89]	+9.36 [+9.13, +9.60]	67.77 [67.38, 68.17]	41.70 [41.30, 42.11]	82.69 [82.24, 83.14]
4	ADAPT	58.42 [58.09, 58.74]	88.20 [87.88, 88.52]	−4.78 [−5.06, −4.52]	+0.72 [+0.43, +1.02]	62.48 [62.13, 62.83]	26.72 [26.45, 26.98]	91.77 [91.42, 92.12]
8	Frozen CLIP	62.11 [61.86, 62.36]	87.76 [87.56, 87.96]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.11 [61.86, 62.36]	35.79 [35.57, 36.00]	90.18 [89.91, 90.43]
8	MOON	71.00 [70.68, 71.31]	81.19 [80.94, 81.43]	−6.57 [−6.73, −6.40]	+15.45 [+15.25, +15.66]	70.22 [69.91, 70.52]	67.56 [67.18, 67.94]	77.35 [77.02, 77.68]
8	StatA	68.12 [67.83, 68.41]	84.62 [84.40, 84.85]	−3.14 [−3.26, −3.01]	+9.14 [+8.97, +9.32]	66.89 [66.61, 67.17]	49.54 [49.23, 49.85]	80.92 [80.61, 81.24]
8	ADAPT	61.74 [61.49, 61.98]	83.69 [83.45, 83.92]	−4.07 [−4.26, −3.89]	+3.70 [+3.52, +3.89]	62.11 [61.86, 62.36]	35.79 [35.57, 36.00]	90.14 [89.88, 90.40]
16	Frozen CLIP	62.15 [61.97, 62.33]	81.65 [81.49, 81.81]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.15 [61.97, 62.33]	48.90 [48.72, 49.07]	87.71 [87.52, 87.90]
16	MOON	67.44 [67.22, 67.65]	74.19 [74.00, 74.38]	−7.46 [−7.59, −7.33]	+12.75 [+12.60, +12.90]	67.32 [67.10, 67.53]	72.37 [72.09, 72.64]	74.61 [74.39, 74.84]
16	StatA	64.96 [64.76, 65.16]	77.31 [77.13, 77.48]	−4.35 [−4.45, −4.24]	+7.15 [+7.03, +7.29]	64.97 [64.77, 65.17]	58.45 [58.23, 58.67]	78.88 [78.65, 79.10]
16	ADAPT	62.10 [61.92, 62.28]	77.48 [77.30, 77.65]	−4.18 [−4.30, −4.05]	+4.13 [+4.01, +4.26]	62.15 [61.97, 62.33]	48.90 [48.72, 49.07]	87.70 [87.51, 87.89]
32	Frozen CLIP	62.18 [62.04, 62.31]	74.71 [74.59, 74.84]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.18 [62.04, 62.31]	65.40 [65.26, 65.53]	84.49 [84.34, 84.63]
32	MOON	61.95 [61.80, 62.11]	66.33 [66.19, 66.47]	−8.39 [−8.49, −8.29]	+8.16 [+8.07, +8.26]	62.75 [62.60, 62.90]	77.94 [77.74, 78.14]	70.54 [70.39, 70.70]
32	StatA	61.04 [60.90, 61.19]	69.25 [69.11, 69.39]	−5.46 [−5.56, −5.37]	+4.33 [+4.24, +4.43]	62.16 [62.01, 62.31]	69.25 [69.09, 69.41]	76.01 [75.85, 76.17]
32	ADAPT	62.23 [62.09, 62.37]	70.65 [70.51, 70.78]	−4.07 [−4.16, −3.97]	+4.12 [+4.03, +4.21]	62.18 [62.04, 62.32]	65.40 [65.26, 65.53]	84.48 [84.34, 84.62]

Table 70: Fixed-$N = 96$ support-sparsity intervention, OpenAI ViT-B/32, skewed class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	62.63 [62.23, 63.02]	92.94 [92.69, 93.19]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.63 [62.23, 63.02]	27.80 [27.52, 28.08]	91.05 [90.70, 91.40]
4	MOON	72.67 [72.19, 73.15]	88.39 [88.03, 88.74]	−4.55 [−4.77, −4.33]	+14.60 [+14.31, +14.88]	71.43 [70.95, 71.90]	62.75 [62.21, 63.28]	79.75 [79.28, 80.22]
4	StatA	69.67 [69.22, 70.12]	91.08 [90.77, 91.39]	−1.87 [−2.01, −1.72]	+8.90 [+8.66, +9.15]	67.65 [67.22, 68.09]	42.26 [41.85, 42.68]	82.55 [82.10, 82.99]
4	ADAPT	57.15 [56.77, 57.51]	88.25 [87.90, 88.59]	−4.69 [−4.97, −4.42]	−0.79 [−1.11, −0.47]	62.63 [62.23, 63.02]	27.80 [27.52, 28.08]	91.03 [90.68, 91.39]
8	Frozen CLIP	62.14 [61.84, 62.44]	87.76 [87.53, 87.99]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.14 [61.84, 62.44]	36.78 [36.55, 37.01]	88.48 [88.21, 88.76]
8	MOON	70.77 [70.40, 71.14]	82.36 [82.06, 82.67]	−5.40 [−5.59, −5.21]	+14.03 [+13.80, +14.25]	69.79 [69.43, 70.15]	66.91 [66.53, 67.28]	75.83 [75.50, 76.17]
8	StatA	68.21 [67.87, 68.56]	85.36 [85.09, 85.63]	−2.41 [−2.55, −2.26]	+8.48 [+8.28, +8.67]	66.72 [66.39, 67.06]	49.45 [49.14, 49.76]	79.95 [79.63, 80.27]
8	ADAPT	59.77 [59.49, 60.05]	83.70 [83.42, 83.97]	−4.07 [−4.27, −3.86]	+1.70 [+1.48, +1.91]	62.14 [61.84, 62.44]	36.79 [36.55, 37.01]	88.46 [88.19, 88.73]
16	Frozen CLIP	62.13 [61.92, 62.34]	81.68 [81.50, 81.87]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.13 [61.92, 62.34]	49.37 [49.18, 49.55]	85.61 [85.40, 85.81]
16	MOON	67.80 [67.54, 68.06]	75.35 [75.12, 75.58]	−6.33 [−6.48, −6.19]	+12.00 [+11.84, +12.16]	67.30 [67.04, 67.55]	72.05 [71.77, 72.33]	71.79 [71.55, 72.03]
16	StatA	65.78 [65.53, 66.02]	78.50 [78.30, 78.71]	−3.18 [−3.30, −3.06]	+6.83 [+6.69, +6.97]	65.16 [64.92, 65.40]	58.45 [58.22, 58.69]	77.07 [76.84, 77.31]
16	ADAPT	61.80 [61.59, 62.01]	77.89 [77.68, 78.10]	−3.79 [−3.93, −3.65]	+3.46 [+3.32, +3.60]	62.14 [61.92, 62.35]	49.38 [49.19, 49.56]	85.60 [85.39, 85.80]
32	Frozen CLIP	62.11 [61.96, 62.27]	74.72 [74.57, 74.86]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	62.11 [61.96, 62.27]	65.22 [65.08, 65.36]	81.68 [81.52, 81.83]
32	MOON	62.69 [62.51, 62.87]	67.33 [67.17, 67.50]	−7.38 [−7.49, −7.27]	+7.96 [+7.86, +8.07]	62.74 [62.57, 62.92]	78.24 [78.02, 78.45]	65.93 [65.76, 66.11]
32	StatA	62.36 [62.19, 62.54]	70.52 [70.36, 70.69]	−4.19 [−4.29, −4.09]	+4.45 [+4.34, +4.55]	62.57 [62.40, 62.74]	69.59 [69.41, 69.75]	73.03 [72.87, 73.20]
32	ADAPT	62.11 [61.96, 62.27]	71.10 [70.95, 71.26]	−3.61 [−3.71, −3.52]	+3.61 [+3.52, +3.71]	62.11 [61.96, 62.27]	65.22 [65.09, 65.37]	81.67 [81.51, 81.82]

Table 71: Fixed-$N = 96$ support-sparsity intervention, OpenAI ViT-B/16, balanced class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	66.04 [65.68, 66.39]	94.47 [94.26, 94.68]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.04 [65.68, 66.39]	31.37 [31.06, 31.69]	91.42 [91.05, 91.78]
4	MOON	74.93 [74.49, 75.38]	89.56 [89.25, 89.86]	−4.92 [−5.12, −4.71]	+13.81 [+13.54, +14.08]	73.93 [73.49, 74.36]	66.23 [65.72, 66.74]	81.86 [81.41, 82.31]
4	StatA	72.42 [72.00, 72.86]	92.36 [92.09, 92.63]	−2.11 [−2.25, −1.97]	+8.50 [+8.27, +8.73]	70.91 [70.51, 71.33]	48.71 [48.29, 49.16]	82.88 [82.43, 83.34]
4	ADAPT	62.11 [61.77, 62.45]	88.90 [88.56, 89.24]	−5.57 [−5.87, −5.28]	+1.64 [+1.33, +1.96]	66.04 [65.69, 66.39]	31.38 [31.07, 31.69]	91.42 [91.05, 91.78]
8	Frozen CLIP	65.59 [65.33, 65.84]	89.94 [89.74, 90.13]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.59 [65.33, 65.84]	40.39 [40.15, 40.64]	89.61 [89.33, 89.88]
8	MOON	73.44 [73.13, 73.76]	83.46 [83.21, 83.71]	−6.47 [−6.64, −6.31]	+14.33 [+14.12, +14.54]	72.66 [72.35, 72.98]	70.86 [70.50, 71.22]	78.59 [78.25, 78.91]
8	StatA	71.14 [70.83, 71.44]	86.48 [86.25, 86.72]	−3.45 [−3.58, −3.33]	+9.00 [+8.82, +9.19]	70.06 [69.77, 70.36]	55.80 [55.46, 56.13]	80.87 [80.54, 81.19]
8	ADAPT	65.19 [64.94, 65.44]	84.91 [84.65, 85.16]	−5.02 [−5.23, −4.83]	+4.63 [+4.43, +4.83]	65.59 [65.34, 65.84]	40.40 [40.15, 40.64]	89.59 [89.31, 89.86]
16	Frozen CLIP	65.54 [65.35, 65.72]	84.32 [84.16, 84.48]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.54 [65.35, 65.72]	53.10 [52.91, 53.29]	87.63 [87.44, 87.82]
16	MOON	70.10 [69.89, 70.32]	77.01 [76.82, 77.19]	−7.31 [−7.44, −7.18]	+11.87 [+11.72, +12.03]	69.90 [69.69, 70.12]	75.09 [74.83, 75.36]	75.77 [75.55, 75.99]
16	StatA	68.31 [68.10, 68.52]	79.64 [79.46, 79.81]	−4.68 [−4.79, −4.57]	+7.45 [+7.32, +7.59]	68.21 [68.01, 68.41]	63.73 [63.48, 63.96]	78.81 [78.58, 79.03]
16	ADAPT	65.50 [65.32, 65.68]	79.97 [79.79, 80.15]	−4.35 [−4.48, −4.22]	+4.31 [+4.18, +4.44]	65.54 [65.35, 65.72]	53.10 [52.91, 53.29]	87.63 [87.43, 87.82]
32	Frozen CLIP	65.59 [65.45, 65.72]	77.82 [77.69, 77.95]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.59 [65.45, 65.72]	68.45 [68.32, 68.59]	84.79 [84.65, 84.93]
32	MOON	64.88 [64.73, 65.03]	69.43 [69.29, 69.58]	−8.39 [−8.49, −8.29]	+7.68 [+7.58, +7.78]	65.60 [65.44, 65.75]	80.20 [80.01, 80.40]	71.74 [71.58, 71.90]
32	StatA	64.57 [64.42, 64.73]	72.31 [72.17, 72.45]	−5.51 [−5.60, −5.42]	+4.50 [+4.40, +4.59]	65.45 [65.30, 65.60]	72.76 [72.59, 72.92]	76.27 [76.10, 76.43]
32	ADAPT	65.66 [65.53, 65.80]	73.88 [73.75, 74.02]	−3.94 [−4.03, −3.85]	+4.01 [+3.92, +4.10]	65.59 [65.45, 65.72]	68.45 [68.32, 68.59]	84.78 [84.64, 84.92]

Table 72: Fixed-$N = 96$ support-sparsity intervention, OpenAI ViT-B/16, skewed class counts; five-dataset macro. Entries are mean [95% CI]. Accuracies and support diagnostics are percentages; Δ_{rank} and Δ_R are percentage points.

K	Method	A_{full}	A_{support}	Δ_{rank}	Δ_R	A_{SO}	P	R
4	Frozen CLIP	66.18 [65.78, 66.58]	94.52 [94.29, 94.75]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	66.18 [65.78, 66.58]	32.73 [32.39, 33.07]	90.78 [90.40, 91.16]
4	MOON	74.71 [74.23, 75.19]	90.02 [89.68, 90.36]	−4.50 [−4.73, −4.28]	+13.03 [+12.76, +13.32]	73.62 [73.15, 74.09]	66.40 [65.88, 66.90]	80.94 [80.49, 81.39]
4	StatA	72.24 [71.78, 72.72]	92.57 [92.28, 92.86]	−1.95 [−2.10, −1.80]	+8.02 [+7.78, +8.26]	70.73 [70.29, 71.19]	49.24 [48.80, 49.69]	82.65 [82.19, 83.11]
4	ADAPT	61.17 [60.78, 61.56]	88.90 [88.52, 89.29]	−5.62 [−5.95, −5.30]	+0.61 [+0.25, +0.97]	66.18 [65.78, 66.58]	32.73 [32.39, 33.07]	90.78 [90.40, 91.16]
8	Frozen CLIP	65.61 [65.31, 65.91]	90.01 [89.78, 90.23]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.61 [65.31, 65.91]	41.43 [41.17, 41.68]	88.04 [87.74, 88.33]
8	MOON	73.32 [72.95, 73.69]	84.61 [84.31, 84.91]	−5.40 [−5.59, −5.21]	+13.10 [+12.88, +13.34]	72.38 [72.02, 72.73]	70.32 [69.95, 70.68]	77.29 [76.96, 77.62]
8	StatA	71.12 [70.76, 71.47]	87.21 [86.94, 87.49]	−2.80 [−2.95, −2.65]	+8.30 [+8.10, +8.50]	69.86 [69.52, 70.20]	55.44 [55.11, 55.77]	80.01 [79.68, 80.34]
8	ADAPT	63.38 [63.09, 63.67]	85.01 [84.71, 85.31]	−5.00 [−5.23, −4.77]	+2.77 [+2.53, +3.02]	65.61 [65.31, 65.91]	41.43 [41.18, 41.69]	88.03 [87.73, 88.31]
16	Frozen CLIP	65.59 [65.38, 65.81]	84.36 [84.18, 84.54]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.59 [65.38, 65.81]	53.64 [53.44, 53.84]	85.71 [85.50, 85.91]
16	MOON	70.50 [70.24, 70.76]	78.20 [77.97, 78.42]	−6.16 [−6.31, −6.02]	+11.07 [+10.90, +11.24]	69.97 [69.71, 70.23]	74.88 [74.60, 75.16]	73.21 [72.98, 73.45]
16	StatA	69.01 [68.76, 69.27]	80.70 [80.48, 80.91]	−3.66 [−3.78, −3.53]	+7.08 [+6.93, +7.22]	68.37 [68.12, 68.62]	63.49 [63.25, 63.73]	77.38 [77.15, 77.61]
16	ADAPT	65.27 [65.06, 65.48]	80.29 [80.08, 80.49]	−4.07 [−4.22, −3.93]	+3.75 [+3.60, +3.90]	65.59 [65.38, 65.81]	53.65 [53.45, 53.85]	85.70 [85.50, 85.91]
32	Frozen CLIP	65.54 [65.38, 65.70]	77.80 [77.65, 77.94]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	65.54 [65.38, 65.70]	68.33 [68.19, 68.47]	82.21 [82.06, 82.36]
32	MOON	65.70 [65.52, 65.88]	70.56 [70.39, 70.72]	−7.24 [−7.35, −7.13]	+7.40 [+7.29, +7.51]	65.63 [65.45, 65.81]	80.27 [80.07, 80.48]	67.70 [67.52, 67.87]
32	StatA	65.81 [65.63, 65.99]	73.44 [73.27, 73.60]	−4.36 [−4.46, −4.26]	+4.63 [+4.53, +4.73]	65.90 [65.73, 66.08]	73.16 [72.98, 73.33]	73.65 [73.48, 73.82]
32	ADAPT	65.56 [65.40, 65.72]	74.16 [74.00, 74.32]	−3.63 [−3.73, −3.54]	+3.65 [+3.56, +3.75]	65.54 [65.38, 65.70]	68.33 [68.19, 68.47]	82.21 [82.06, 82.36]