When Catastrophic Inheritance Meets Forgetting in Continual Adaptation of Foundation Models

Quanyu Zhang[1,2], **Zhongyi Han**[1,2,*], **Zhenxue Chen**[3],
Xiao-Long Yin[1,*], **Henggui Zhang**[2], **Shuo Li**[4]
[1]School of Software, Shandong University [2]Beijing Academy of Artificial Intelligence
[3]School of Control Science and Engineering, Shandong University
[4]Case Western Reserve University
qyzhang@mail.sdu.edu.cn zhongyi.han@sdu.edu.cn chenzhenxue@sdu.edu.cn
yinxl@lamda.nju.edu.cn H.zhang-3@manchester.ac.uk slishuo@gmail.com
[*]Corresponding authors

Abstract

Pretraining noise in foundation models has been shown to significantly impair the generalization performance of downstream tasks, leading to catastrophic inheritance. However, when downstream tasks arrive sequentially, how pretraining noise affects continual adaptation, which involves the forgetting of previously learned knowledge, remains underexplored. This paper is the first to study this problem. Using asymmetric label noise as a realistic setting, we find that pretraining noise persistently degrades performance on new tasks during continual adaptation and may further increase forgetting of previously adapted tasks in some settings. Further analysis shows that the persistent effect of pretraining noise mainly manifests as inherited output contraction, where the logit gap between the true class and competing classes is compressed. To mitigate this issue, we propose Ambiguous Boundary Correction (ABC), a method that performs anchor-guided neighborhood reweighting to correct ambiguous regions induced by inherited output contraction. Experiments on synthetic and real-world noisy pretrained models show that ABC improves robustness to pretraining noise under both domain shifts and class shifts.

1 Introduction

Foundation models [Bommasani et al., 2021, Awais et al., 2025, Dong et al., 2026, Chen et al., 2026] have demonstrated remarkable transferability and generalization through large-scale pretraining. However, as datasets continue to scale, defects such as annotation errors [Liu et al., 2025], image–text mismatches [Duan et al., 2024, Cheng et al., 2025], sampling bias [Parashar et al., 2024, Chen et al., 2025b], and other data quality issues [Nguyen et al., 2022, Lee et al., 2022, Subramanyam et al., 2025, Kang et al., 2025, Ghosh et al., 2025, Wei et al., 2025] become increasingly difficult to eliminate. Recent studies have shown that such pretraining defects can be internalized by models and subsequently inherited by downstream tasks, leading to degraded prediction performance and impaired generalization [Chen et al., 2024b, Chang et al., 2024]. This phenomenon, termed catastrophic inheritance [Chen et al., 2024a], has emerged as an important challenge in the development of foundation models, with pretraining noise being one of its prevalent and representative sources.

Meanwhile, foundation models often need to be continually adapted to dynamic downstream environments. As data distributions shift, application scenarios diversify, and task requirements evolve, models are expected to acquire new knowledge over time. However, when learning new tasks or adapting to new data distributions sequentially, updates driven by newly observed data can interfere with previously acquired knowledge, causing performance on earlier tasks to decline. This phe-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

nomenon, known as catastrophic forgetting [Wang et al., 2024a, Shi et al., 2025, Li et al., 2025b, Luo et al., 2025], is a fundamental challenge in continual adaptation.

What happens when catastrophic inheritance meets forgetting during the continual adaptation of foundation models? As the model sequentially adapts to new tasks, will the defects inherited from pretraining be gradually corrected by downstream supervision, or will they persist and further affect the learning and retention of downstream tasks? Understanding this interaction is important for revealing how pretraining defects influence foundation models under continual adaptation.

To answer this question, we conduct the first systematic study of how pretraining noise affects continual adaptation. We construct a controlled setting by introducing different levels of asymmetric label noise [Song et al., 2025, Dix et al., 2025] during ImageNet-1K [Russakovsky et al., 2015] pretraining and sequentially adapting the resulting Vision Transformer (ViT) models [Dosovitskiy et al., 2021] to downstream domains [Zhou et al., 2025b]. In this setting, we observe a distinct pattern: while previously learned tasks are progressively forgotten, the degradation inherited from noisy pretraining does not gradually disappear. Instead, it persists throughout the task sequence, reducing performance on newly arriving tasks and, in some settings, further increasing forgetting of previously adapted tasks (see Figure 1). Further analysis shows that this persistent effect is associated with inherited output contraction, where noisy pretraining compresses the logit gap between the true class and competing classes, resulting in lower confidence and smaller representation margins.

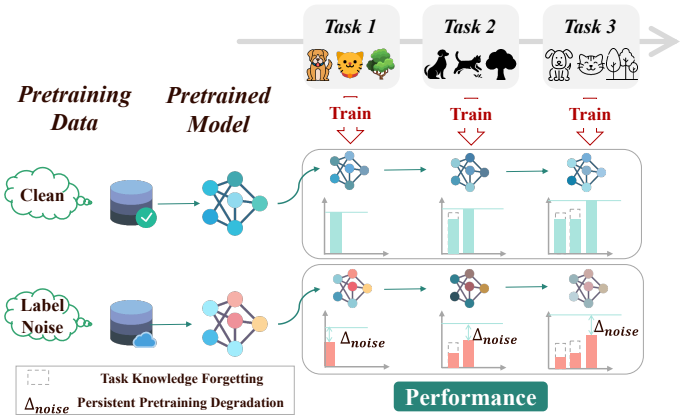


Figure 1: Illustration of the impact of pretraining noise on continual adaptation. The performance gap between clean and noisy pretraining on the current task reflects the persistent degradation in new-task learning caused by pretraining noise. The decline in previously learned tasks during subsequent adaptation reflects catastrophic forgetting, while the extent to which pretraining noise further increases this forgetting varies across settings.

Building on these findings, we propose Ambiguous Boundary Correction (ABC), which converts inherited output contraction into a trainable correction mechanism. Specifically, ABC identifies low-confidence predictions as indicators of contracted regions and propagates corrective supervision to their cross-class neighbors. By reweighting these cross-class neighboring samples, ABC focuses adaptation on ambiguous regions most affected by inherited output contraction. Experiments on both synthetic and real-world noisy pretrained models demonstrate that ABC improves robustness to pretraining noise across domain and class shifts.

The main contributions of this paper are as follows:

- **Influence.** We conduct the first systematic study of the impact of pretraining noise on continual adaptation and find that it persistently degrades performance on new tasks and may further increase forgetting of previously adapted tasks in some cases.
- **Analysis.** We identify inherited output contraction as a key mechanism of catastrophic inheritance, showing how pretraining noise induces persistent bias.
- **Mitigation.** We propose ABC, a mechanism-guided method that mitigates inherited bias through anchor-guided neighborhood reweighting, without requiring access to pretraining.

2 Related Work

Catastrophic inheritance. As pretraining datasets continue to scale, data noise and bias become increasingly difficult to eliminate, including issues such as label contamination [Khanal et al., 2024] and image–text mismatches [Wang et al., 2023]. These imperfections can undermine the

generalization ability and reliability [Li et al., 2025a, 2026] of foundation models. Existing approaches primarily address pretraining data issues from two directions. One line of work seeks to improve data quality through cleaning or filtering, but is often constrained by substantial computational cost [Radenovic et al., 2023, Chen et al., 2023, Caron et al., 2024, Vasa et al., 2025]. Another line of work focuses on improving the pretraining process through noise-robust optimization strategies [Tao et al., 2025, Lu et al., 2025], such as RobustFormer [Bastola et al., 2026] and MIRAGE [Cheng et al., 2025]. However, these methods typically require access to pretraining data or modifications to the pretraining pipeline. More recently, several studies have begun to investigate how pretraining noise transfers to downstream tasks. Chen et al. [2025a] reported a negative correlation between pretraining noise and downstream robustness, and similar findings have also been observed in medical applications [Sun et al., 2025]. Nevertheless, these studies focus on single-stage downstream adaptation and do not examine how pretraining noise influences continual adaptation over time.

Catastrophic forgetting. Catastrophic forgetting is a fundamental challenge in continual adaptation, referring to the loss of previously acquired knowledge as a model learns new tasks or adapts to new data distributions [Parisi et al., 2019, Liu et al., 2023]. This problem is commonly studied under sequential adaptation settings, which are typically categorized into task shifts, domain shifts, and class shifts according to the availability of task identity and changes in the label space [Van de Ven et al., 2022]. Task shifts assume that task identity is available during inference [Oren and Wolf, 2021]. Domain shifts consider settings where the label space remains unchanged while the input distribution changes over time [Wang et al., 2024b, Liu et al., 2024]. Class shifts address scenarios in which the label space gradually expands and task identity is unavailable at test time [Masana et al., 2022]. With the rise of foundation models, recent studies have explored how to leverage the strong transferability of pretrained models for continual adaptation [Du et al., 2022, Zhou et al., 2024, Wang et al., 2024a], often through parameter-efficient tuning methods such as LoRA for downstream adaptation [Wang et al., 2025a, Coleman et al., 2025]. However, these methods typically assume that pretrained models provide reliable and unbiased initial representations.

3 How Pretraining Noise Affects Continual Adaptation

In this section, we use controlled experiments to quantify the impact of pretraining noise on continual adaptation. We find that pretraining noise has a persistent effect across the task sequence, consistently weakening learning on new tasks and, in some cases, increasing forgetting of previously adapted tasks. Specifically, we inject different levels of asymmetric label noise during pretraining and evaluate the resulting models under continual adaptation with domain shifts.

3.1 Experimental Setup

Pretraining noise. We use ImageNet-1K [Russakovsky et al., 2015] as the pretraining dataset. ImageNet-1K categories are associated with WordNet synsets, which naturally provide a semantic hierarchy. To simulate realistic semantic annotation errors in pretraining data, we introduce asymmetric label noise according to WordNet semantic similarity. Specifically, for samples selected to be corrupted, their labels are more likely to be replaced by labels from sibling categories that share the same parent node in the WordNet hierarchy. This design reflects the fact that real-world annotation errors are often semantically biased: for example, "ferret" is more likely to be confused with "mink" or "polecat" than with unrelated categories such as "car" or "chair". We use five pretraining noise ratios, $\rho \in \{0\%, 5\%, 10\%, 20\%, 30\%\}$.

Pretraining protocol. All models use the ViT-B/16 architecture [Dosovitskiy et al., 2021] with an input resolution of 224×224. The training setup follows the DeiT-style configuration [Touvron et al., 2022]. We use the LAMB optimizer [You et al., 2019] with an initial learning rate of 5×10^{-4}, weight decay of 0.02, and batch size of 512. The models are trained for 300 epochs with 5 warm-up epochs. We use BCE loss, stochastic depth with a drop rate of 0.2 [Huang et al., 2016], repeated augmentation [Hoffer et al., 2019], LayerScale [Touvron et al., 2021], and gradient clipping with a threshold of 1.0. All hyperparameters are kept the same across different noise ratios.

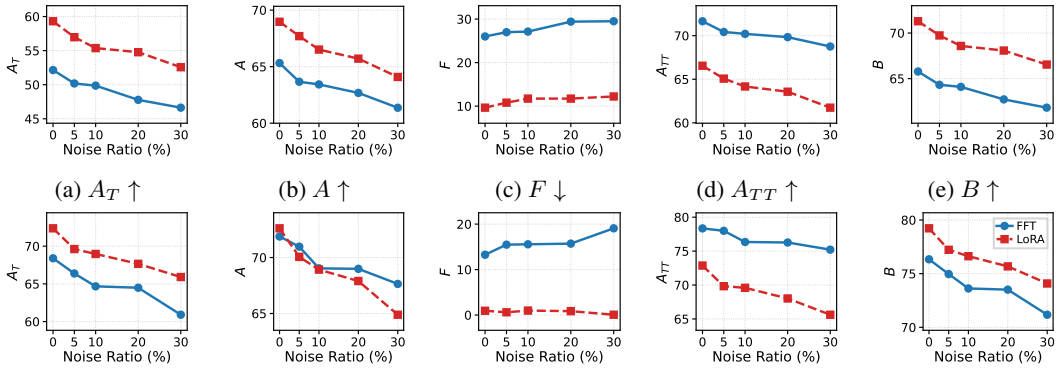
Downstream tasks. We evaluate downstream performance under domain shifts. Two multi-domain datasets are used: DomainNet [Peng et al., 2019] and OfficeHome [Venkateswara et al., 2017]. DomainNet contains 345 categories across multiple visual domains; we use four domains in our experiments: Real World, Clipart, Painting, and Sketch. OfficeHome contains 65 categories from

four domains: Art, Clipart, Product, and Real World. Each domain is treated as a new task. After learning a task, the model cannot access the training data of previous tasks.

Fine-tuning strategies. We consider two downstream adaptation strategies: full fine-tuning (FFT) and LoRA. In FFT, all model parameters are updated for each new task. In LoRA [Hu et al., 2022], the pretrained backbone is kept frozen, and only low-rank adaptation parameters inserted into the fully connected layers of the MLP blocks are optimized. For both strategies, we use SGD [Bottou, 2010] with a learning rate of 0.001, train each task for 15 epochs, and use a batch size of 128.

Evaluation metrics. Let there be T tasks, and let $a_{i,j}$ denote the test accuracy on the j-th task after the model has completed learning the i-th task, where $1 \leq j \leq i \leq T$. In addition to the commonly used metrics A_T, A, and F whose definitions are provided in Appendix A.1, we introduce two additional metrics to better measure learning ability on new tasks and overall performance during continual adaptation:

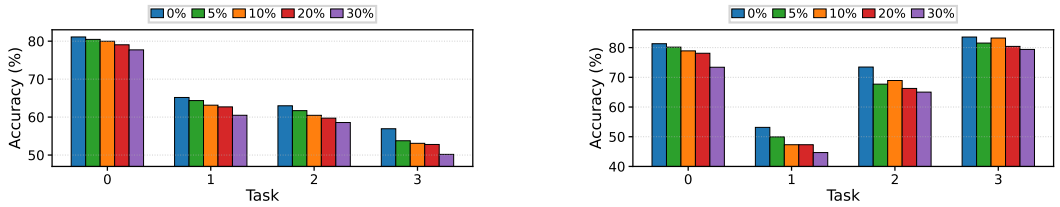
- **Current-task Average Accuracy**: $A_{TT} = \frac{1}{T}\sum_{i=1}^{T} a_{i,i}$. It measures the model's ability to learn each newly arriving task.
- **Comprehensive Metric**: $B = \frac{1}{4}(A_T + A + (100 - F) + A_{TT})$. It provides a unified summary of final performance, historical performance, new-task learning, and forgetting. Since all components are measured on the same percentage scale, we assign them equal weights to obtain an overall evaluation score.



(a) $A_T \uparrow$ (b) $A \uparrow$ (c) $F \downarrow$ (d) $A_{TT} \uparrow$ (e) $B \uparrow$

(f) $A_T \uparrow$ (g) $A \uparrow$ (h) $F \downarrow$ (i) $A_{TT} \uparrow$ (j) $B \uparrow$

Figure 2: Comparison of FFT and LoRA under increasing noise levels on DomainNet (first row) and OfficeHome (second row) across five metrics: A_T, A, F, A_{TT}, B. Higher noise levels generally correspond to lower accuracy-related metrics and higher forgetting rates.



(a) DomainNet, LoRA (b) OfficeHome, LoRA

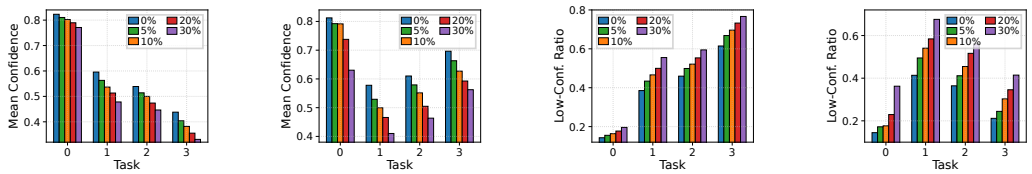
Figure 3: Current-task accuracy $a_{i,i}$ of LoRA under different noise levels on DomainNet and OfficeHome. As the noise ratio increases, the accuracy on new tasks gradually decreases, and the magnitude of this decline remains relatively stable as more tasks are learned.

3.2 Observation

Based on the experimental results, we obtain two consistent observations:

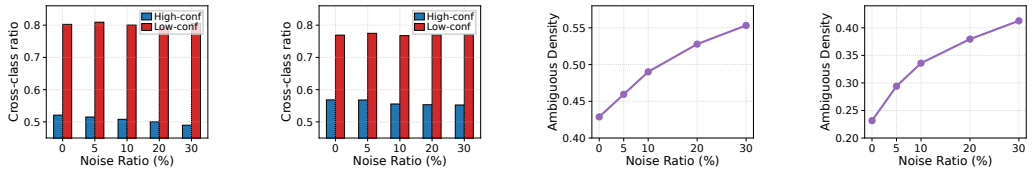
Observation 1: Pretraining noise has a persistent impact on downstream continual adaptation. Figure 2 compares FFT and LoRA under different noise ratios. As the noise level increases, A_T, A, A_{TT}, and B consistently decline across both datasets, with no sign of saturation. For instance, under LoRA on OfficeHome, A_{TT} decreases by 7.25 points as the noise ratio increases from 0% to 30%. Figure 3 further shows that the current-task accuracy $a_{i,i}$ remains consistently degraded throughout the task sequence, indicating that the impact of pretraining noise on newly encountered tasks remains relatively stable and does not weaken as the number of tasks increases.

Observation 2: Pretraining noise can increase forgetting of previously adapted tasks in some settings. Figure 2 shows that pretraining noise can further increase the forgetting rate F in some settings. Compared with its effect on new-task learning, the impact on forgetting is more complex, as forgetting is jointly influenced by factors such as initial task-learning performance, interference from subsequent tasks, and model adaptation capacity. For example, under FFT on OfficeHome, F increases by 5.81 points when the noise ratio rises from 0% to 30%. This suggests that pretraining noise may also impair the model's ability to retain knowledge from previous tasks.



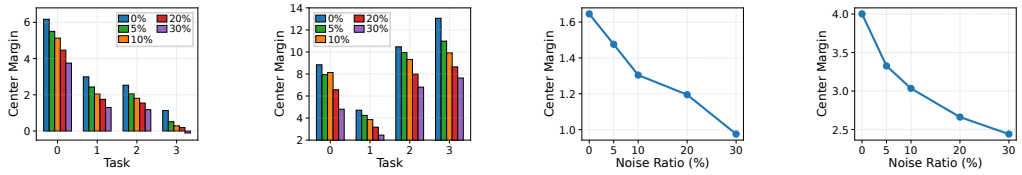
(a) DomainNet, Mean (b) OfficeHome, Mean (c) DomainNet, Low (d) OfficeHome, Low

Figure 4: Prediction confidence under different noise ratios. The figure shows the mean confidence and the ratio of low-confidence samples (below 0.5) on DomainNet and OfficeHome. A higher noise ratio generally corresponds to lower mean confidence and a higher low-confidence ratio.



DomainNet (a) Cross-class ratio | OfficeHome (b) Cross-class ratio | DomainNet (c) Ambiguous density | OfficeHome (d) Ambiguous density

Figure 5: Local structure of low-confidence samples under different noise ratios. We first compute the cross-class proportion among each sample's top-5 nearest neighbors in the feature space. The first two subfigures compare this proportion between low-confidence and high-confidence samples. The last two subfigures report ambiguous density, defined as the product of the low-confidence sample ratio and its corresponding cross-class proportion.



(a) DomainNet, Current (b) OfficeHome, Current (c) DomainNet, Final (d) OfficeHome, Final

Figure 6: Representation margin under different noise ratios. The figure shows the current-task margin and final margin on DomainNet and OfficeHome, where the margin is defined as the difference between the distance to the nearest negative-class center and the distance to the same-class center. A higher noise ratio generally corresponds to a smaller margin.

3.3 Inherited Output Contraction

The two observations above indicate that pretraining noise changes the output state used during continual adaptation, leading to degraded learning on new tasks and weakened retention of previously adapted tasks. Let z_i denote the logits of sample x_i and y_i its ground-truth label. We define the true-class logit gap as

$$g_i(\theta) = z_{i,y_i} - \max_{k \neq y_i} z_{i,k}. \tag{1}$$

A larger $g_i(\theta)$ generally indicates a stronger true-class advantage and higher prediction confidence, whereas a smaller gap flattens the predictive distribution and lowers confidence.

Pretraining noise compresses this margin. Clean supervision provides consistent class directions for semantically similar samples, thereby increasing separation between the true-class logit and competing logits. In contrast, noisy supervision introduces conflicting labels within local semantic regions, causing neighboring samples to simultaneously reinforce different class logits and resulting in a more compact output distribution. Under this local competition mechanism, the output space gradually enters a low-margin stable state. Figure 4 shows that as the noise ratio increases, the mean confidence decreases while the proportion of low-confidence samples increases, indicating that this low-margin state persistently affects learning across the task sequence.

This contraction is more pronounced in regions where classes are difficult to separate. Figure 5 shows that low-confidence samples exhibit higher cross-class neighborhood overlap, indicating that contracted outputs concentrate in locally class-conflicting regions. Figure 6 further shows that increasing pretraining noise reduces the class margin, suggesting that output contraction is accompanied by compressed class separation in the representation space.

Output contraction also weakens the retention of previously adapted tasks. For adapted samples, a larger logit gap provides a wider safety margin against subsequent updates, whereas a smaller gap makes them more susceptible to prediction drift or misclassification during later adaptation.

These findings suggest that the main effect of pretraining noise in continual adaptation is not only a general performance drop, but a persistent contraction of output margins around ambiguous cross-class regions. More details are provided in Appendix A.2. This observation directly motivates a training strategy that identifies such regions and strengthens supervision on them.

4 Ambiguous Boundary Correction for Inherited Output Contraction

Motivated by the inherited output contraction observed in Section 3, we propose Ambiguous Boundary Correction (ABC), a mechanism-guided boundary correction method (see Figure 7). Since inherited contraction mainly appears around cross-class ambiguous regions, uniform adaptation cannot explicitly repair these ambiguous regions. The key idea of ABC is to use low-confidence samples as observable indicators of contracted output regions and then identify cross-class neighbors around them as ambiguous boundary samples for reweighted training.

4.1 Method

Let a mini-batch from the current task be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where z_i denotes logits and f_i denotes normalized penultimate-layer features. Standard adaptation minimizes the cross-entropy loss:

$$L_{\mathrm{CE}} = -\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(y_i|x_i). \tag{2}$$

ABC departs from generic hard-sample reweighting by using low-confidence samples as anchors to locate inherited ambiguous boundaries, rather than simply up-weighting uncertain samples themselves. We define a low-confidence set:

$$c_i = \max_k p_\theta(k|x_i), \quad \mathcal{A} = \{i \mid c_i < \tau\}. \tag{3}$$

For each $i \in \mathcal{A}$, we construct a local neighborhood in feature space and extract cross-class neighbors:

$$\mathcal{S}_i = \mathrm{TopK}_{j \neq i} \big(\cos(f_i, f_j), K \big), \quad \mathcal{N}_i = \{j \in \mathcal{S}_i \mid y_j \neq y_i\}, \tag{4}$$

with $\mathcal{N} = \bigcup_{i \in \mathcal{A}} \mathcal{N}_i$. Then ABC reweights the cross-entropy loss over samples in $\mathcal{N}$:

$$\mathcal{L}_{\text{ABC}} = -\frac{1}{n} \sum_{i=1}^{n} w_i \log p_\theta(y_i|x_i), \tag{5}$$

where

$$w_i = \begin{cases} 1 + \lambda \frac{n}{|\mathcal{N}|}, & \text{if } i \in \mathcal{N}, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

If no sample is selected in $\mathcal{N}$, we set $w_i = 1$ for all samples. The normalization term $\frac{n}{|\mathcal{N}|}$ prevents the additional supervision strength from being affected by the number of selected samples. Therefore, λ controls the relative emphasis on ambiguous cross-class regions.

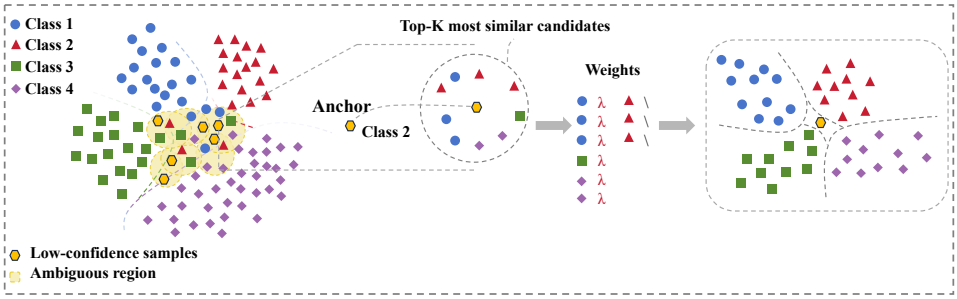
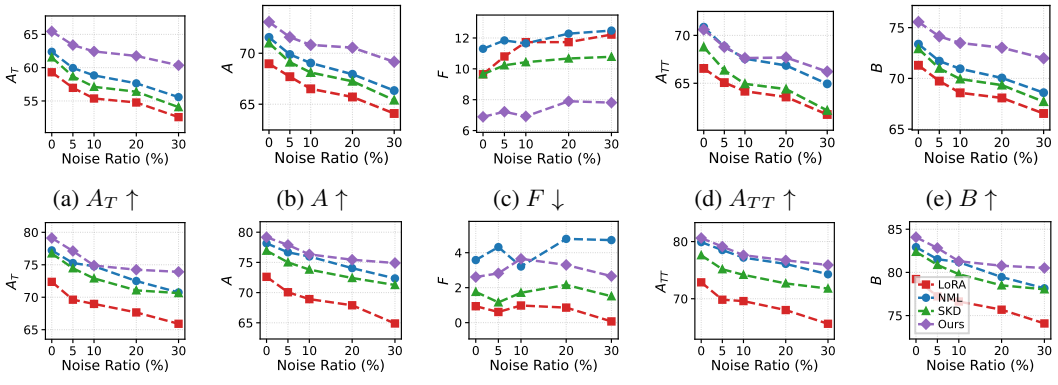


Figure 7: Illustration of Ambiguous Boundary Correction (ABC). Low-confidence samples serve as anchors to identify uncertain boundary regions. ABC retrieves cross-class neighbors around these anchors and assigns higher training weights, encouraging the model to focus on ambiguous regions.

This weighting induces a non-uniform scaling of empirical risk concentrated on samples with high local feature ambiguity. Since such samples concentrate near decision boundaries, the resulting optimization corresponds to an implicit, data-dependent margin regularization that emphasizes regions with small inter-class separation in feature space.



(a) $A_T \uparrow$ (b) $A \uparrow$ (c) $F \downarrow$ (d) $A_{TT} \uparrow$ (e) $B \uparrow$

(f) $A_T \uparrow$ (g) $A \uparrow$ (h) $F \downarrow$ (i) $A_{TT} \uparrow$ (j) $B \uparrow$

Figure 8: Performance comparison under different pretraining noise ratios. Results are reported for LoRA, NML, SKD, and ABC on DomainNet (first row) and OfficeHome (second row) across five evaluation metrics. ABC consistently achieves stronger overall performance and effectively reduces performance degradation as the pretraining noise ratio increases.

4.2 Results

To validate the effectiveness of ABC, we compare it with LoRA, NML [Chen et al., 2025a], and SKD [Sun et al., 2025] under synthetic pretraining noise settings. Both NML and SKD are originally designed with a frozen backbone and incorporate additional adaptation modules or regularization

terms. For a fair comparison, we standardize the backbone adaptation by applying LoRA to all methods, while preserving the original auxiliary components of NML and SKD.

Figure 8 reports performance under varying pretraining noise ratios. ABC achieves the best overall metric B across all noise settings on both datasets, and consistently outperforms baselines in A_T, A, and A_{TT}, indicating improved final performance, adaptation ability, and current-task learning. For the forgetting rate F, ABC also achieves lower values on DomainNet. Notably, ABC exhibits a smaller performance drop as noise ratio increases. In terms of metric B, LoRA decreases from 71.29% to 66.54% on DomainNet, whereas ABC decreases from 75.56% to 71.98%. On OfficeHome, LoRA drops from 79.23% to 74.09%, while ABC drops from 84.08% to 80.51%. These results demonstrate that ABC effectively mitigates the negative impact of pretraining noise.

5 Experiments

To evaluate the effectiveness of ABC under more realistic pretraining conditions, we conduct additional experiments using a web-scale pretrained CLIP/ViT-L/14 model [Cherti et al., 2023], whose pretraining regime naturally involves data imperfections in large-scale image-text pairs. We focus primarily on domain-shift settings and further extend the evaluation to class-shift scenarios, aiming to assess the robustness and generalization ability of ABC.

Table 1: Comparison of different methods across four domain-shift benchmarks. For each metric, ↑ indicates higher is better, ↓ indicates lower is better. Best results are in **bold**, second-best are underlined. Gray rows highlight our methods (Ours and Ours with Feature Projection).

Dataset	Method	$A_T \uparrow$	$A \uparrow$	$F \downarrow$	$A_{TT} \uparrow$	$B \uparrow$
DomainNet	LoRA	80.93	85.60	3.45	83.52	86.65
	NML	81.11	86.58	5.17	84.43	86.73
	SKD	80.37	85.57	3.36	82.89	86.37
	Ours	**82.30**	**86.61**	**2.94**	84.50	**87.62**
	Ours (Feature Projection)	81.15	86.15	4.74	**84.70**	86.82
OfficeHome	LoRA	88.05	87.41	**0.00**	86.82	90.57
	NML	88.27	89.06	1.88	89.69	91.28
	SKD	88.32	88.08	0.42	88.49	91.12
	Ours	**90.97**	**91.19**	1.24	**91.61**	**93.13**
	Ours (Feature Projection)	89.92	90.39	1.86	91.31	92.44
CDDB	LoRA	65.67	61.63	8.64	69.33	72.00
	NML	63.82	61.41	8.94	69.63	71.48
	SKD	59.06	58.58	**0.74**	58.81	68.93
	Ours	**70.51**	**66.09**	8.03	72.45	**75.25**
	Ours (Feature Projection)	68.53	64.78	11.33	**74.22**	74.05
TerraIncognita	LoRA	62.01	70.40	**6.03**	66.28	73.16
	NML	57.56	69.47	17.43	70.63	70.06
	SKD	56.31	67.98	8.36	62.34	69.57
	Ours	65.40	73.64	10.37	72.80	75.37
	Ours (Feature Projection)	**69.26**	**76.84**	13.88	**79.67**	**77.97**

Experimental setup. All experiments are conducted on a computing server equipped with four NVIDIA RTX 4090 GPUs. Model training uses a distributed data-parallel strategy, with a batch size of 128. The optimizer is SGD with a learning rate set to 0.001, and each task is trained for 15 epochs. To ensure a fair and consistent comparison with methods such as NML and SKD, which introduce feature projection modules, we further extend our method with a lightweight **feature projection** module in the real-world noisy pretrained model experiments. This module is used to enhance the adaptability of the feature space, but does not change the core optimization objective or continual adaptation strategy of this paper. Model evaluation uses the five metrics mentioned earlier.

Domain shifts. We evaluate on four representative multi-domain benchmarks: DomainNet, OfficeHome, TerraIncognita [Beery et al., 2018], and CDDB [Li et al., 2023]. DomainNet and OfficeHome

follow the setting in Section 3.1. For TerraIncognita, we use four domains: Location38, Location43, Location46, and Location100. For CDDB, we use five domains: san, whichfaceisreal, biggan, wild, and gaugan. Under domain shifts, each domain is treated as a sequential task, and previous training data are not revisited during adaptation.

Table 2: Comparison of different methods on class-shift benchmarks. For each metric, ↑ indicates higher is better, while ↓ indicates lower is better. Best results are in **bold**, and second-best results are underlined. Gray rows highlight our methods (Ours and Ours with Feature Projection).

Dataset	Method	$A_T \uparrow$	$A \uparrow$	$F \downarrow$	$A_{TT} \uparrow$	$B \uparrow$
CUB-40	LoRA	57.22	66.16	8.64	63.80	69.64
	NML	72.74	76.99	13.32	83.40	79.95
	SKD	67.93	75.75	7.88	74.20	77.50
	Ours	**76.39**	**80.79**	**7.54**	82.42	**83.02**
	Ours (Feature Projection)	76.06	80.07	11.37	**85.16**	82.48
CUB-20	LoRA	47.08	59.71	11.32	57.17	63.16
	NML	66.22	74.23	16.98	81.51	76.25
	SKD	54.88	69.95	**9.38**	63.32	69.69
	Ours	67.19	75.29	10.77	76.84	77.14
	Ours (Feature Projection)	**68.82**	**76.67**	16.87	**84.00**	**78.16**

Results under domain shifts. Table 1 reports performance under real-world noise across all datasets. Our method achieves consistently strong results, obtaining either the best or second-best performance across most metrics. Notably, even with the additional feature projection module, the performance remains stable, indicating that ABC effectively mitigates the adverse impact of real-world pretraining noise while maintaining structural compatibility with existing adaptation modules.

Class shifts. We further evaluate ABC under class-shift settings using CUB-200-2011 [Wah et al., 2011], a fine-grained benchmark containing 200 bird species. In this setting, classes are partitioned into disjoint tasks, and task identity is not available during inference. To comprehensively evaluate different levels of adaptation difficulty, we construct two configurations: CUB-40 (5 tasks with 40 classes each) and CUB-20 (10 tasks with 20 classes each), corresponding to different levels of task granularity and adaptation complexity.

Results under class shifts. Table 2 presents results under class-shift settings. Our method maintains strong performance across both task granularities, consistently outperforming or matching baselines. These results further demonstrate the robustness and generalization ability of the proposed method in task-identity-agnostic continual adaptation scenarios.

Combination with existing continual adaptation methods. We further evaluate the compatibility of our method with existing continual adaptation approaches by integrating it with APER-Adapter [Zhou et al., 2025a] and TUNA [Wang et al., 2025b]. Experiments are conducted using CLIP/ViT-B/16 on DomainNet and OfficeHome under domain-shift settings. The results in Table 3 show that ABC can be incorporated into existing continual adaptation methods and provides consistent gains in several settings. For example, on OfficeHome, integrating our method with TUNA improves B from 89.37% to 91.26%, demonstrating the effectiveness of the proposed approach in enhancing existing methods.

6 Conclusion

This work is the first to characterize the behavior of pretraining noise during downstream continual adaptation, showing that it persistently degrades performance on new tasks across the task sequence and can further increase forgetting in some settings. Our analysis shows that this persistent effect mainly manifests as inherited output contraction: pretraining noise compresses the logit gap between the true class and competing classes, leading to lower confidence, more low-confidence samples, and smaller representation margins during continual adaptation. Based on this finding, we propose Ambiguous Boundary Correction (ABC), a mechanism-guided correction method that mitigates inherited output contraction through anchor-guided neighborhood reweighting. Experiments under synthetic and real-world noise settings show that ABC improves model robustness and generalizes well to both domain-shift and class-shift scenarios.

Table 3: Combination results of our method with existing continual adaptation methods using CLIP/ViT-B/16. For each metric, ↑ indicates higher is better, while ↓ indicates lower is better.

Dataset	Method	$A_T \uparrow$	$A \uparrow$	$F \downarrow$	$A_{TT} \uparrow$	$B \uparrow$
DomainNet	APER-Adapter	77.22	81.52	0.25	77.41	83.97
	APER-Adapter + Ours	77.22	81.56	0.24	77.40	83.98
	TUNA	78.35	83.10	1.69	79.62	84.84
	TUNA + Ours	79.63	84.55	2.26	81.32	85.81
OfficeHome	APER-Adapter	86.02	86.47	0.27	86.15	89.59
	APER-Adapter + Ours	86.97	87.50	0.30	87.14	90.33
	TUNA	86.46	86.01	0.31	85.33	89.37
	TUNA + Ours	88.39	88.62	0.27	88.32	91.26

Acknowledgements

This work was supported by the Key Research and Development Program of Shandong Province (Grant No. 2026CXGC010301), the National Natural Science Foundation of China (Grant No. 62606284 and U23A20389), the Excellent Young Scientists Fund Program (Overseas) of the Shandong Provincial Natural Science Foundation (Grant No. 2026HWYQ-007), the Youth Innovation Science and Technology Support Program for Higher Education Institutions of Shandong Province (Grant No. 2025KJH105), and the Beijing Academy of Artificial Intelligence (BAAI).

References

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025.

Ashish Bastola, Nishant Luitel, Hao Wang, Danda Pani Paudel, Roshni Poudel, and Abolfazl Razi. Robustformer: Noise-robust pre-training for images and videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2284–2294, 2026.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

Mathilde Caron, Alireza Fathi, Cordelia Schmid, and Ahmet Iscen. Web-scale visual entity recognition: An llm-driven data approach. *Advances in Neural Information Processing Systems*, 37:34533–34560, 2024.

Yupeng Chang, Yi Chang, and Yuan Wu. Ba-lora: Bias-alleviating low-rank adaptation to mitigate catastrophic inheritance in large language models. *arXiv preprint arXiv:2408.04556*, 2024.

Chi Chen, Peng Li, Maosong Sun, and Yang Liu. Weakly supervised vision-and-language pre-training with relative representations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8341–8355, 2023.

H Chen, Z Wang, R Tao, H Wei, X Xie, M Sugiyama, B Raj, and J Wang. Impact of noisy supervision in foundation model learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7): 5690–5707, 2025a.

Hao Chen, Bhiksha Raj, Xing Xie, and Jindong Wang. On catastrophic inheritance of large foundation models. *Journal of Data-centric Machine Learning Research*, 2024a.

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Understanding and mitigating the label noise in pre-training on downstream tasks. In *The Twelfth International Conference on Learning Representations*, 2024b.

Haolong Chen, Hanzhi Chen, Zijian Zhao, Kaifeng Han, Guangxu Zhu, Yichen Zhao, Ying Du, Wei Xu, and Qingjiang Shi. An overview of domain-specific foundation model: key technologies, applications and challenges. *Science China Information Sciences*, 69(1):111301, 2026.

Jiahao Chen, Bin Qin, Jiangmeng Li, Hao Chen, and Bing Su. Rethinking the bias of foundation model under long-tailed distribution. *arXiv preprint arXiv:2501.15955*, 2025b.

Pujin Cheng, Yijin Huang, Li Lin, Junyan Lyu, Kenneth Kin-Yip Wong, and Xiaoying Tang. Mirage: Medical image-text pre-training for robustness against noisy environments. *Medical Image Analysis*, page 103912, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

Eric Nuertey Coleman, Luigi Quarantiello, Ziyue Liu, Qinwen Yang, Samrat Mukherjee, Julio Hurtado, and Vincenzo Lomonaco. Parameter-efficient continual fine-tuning: A survey. *arXiv preprint arXiv:2504.13822*, 2025.

Marcel Dix, Gianluca Manca, and Alexander Fay. Measuring the robustness of supervised ml models to label noise in industrial data. In *2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS)*, pages 1–4. IEEE, 2025.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5672–5691, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

Yue Duan, Zhangxuan Gu, Zhenzhe Ying, Lei Qi, Changhua Meng, and Yinghuan Shi. Pc2: Pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 9397–9406, 2024.

Adhiraj Ghosh, Vishaal Udandarao, Thao Nguyen, Matteo Farina, Mehdi Cherti, Jenia Jitsev, Sewoong Oh, Elisa Ricci, Ludwig Schmidt, and Matthias Bethge. Concept-aware batch sampling improves language-image pretraining. *arXiv preprint arXiv:2511.20643*, 2025.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations(ICLR)*, 1(2):3, 2022.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10750–10769, 2025.

Bidur Khanal, Binod Bhattarai, Bishesh Khanal, and Cristian Linte. How does self-supervised pretraining improve robustness against noisy labels across various medical image classification datasets? *arXiv preprint arXiv:2401.07990*, 2024.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, 2022.

Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1339–1349, 2023.

Wenhao Li, Qiangchang Wang, Xianjing Meng, Zhibin Wu, and Yilong Yin. Vt-fsl: Bridging vision and text with llms for few-shot learning. In *NeurIPS*, volume 38, pages 113697–113727, 2025a.

Wenhao Li, Xianjing Meng, Qiangchang Wang, Zhongyi Han, Zhibin Wu, and Yilong Yin. Dvla-rl: Dual-level vision-language alignment with reinforcement learning gating for few-shot learning. In *ICLR*, 2026.

Xinlong Li, Weijieying Ren, Wei Qin, Lei Wang, Tianxiang Zhao, and Richang Hong. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025b.

Chenying Liu, Conrad M Albrecht, Yi Wang, and Xiao Xiang Zhu. Cromss: Cross-modal pretraining with noisy labels for remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Chong Liu, Yi Wang, Dong Li, and Xizhao Wang. Domain-incremental learning without forgetting based on random vector functional link networks. *Pattern recognition*, 151:110430, 2024.

Hao Liu, Yong Zhou, Bing Liu, Jiaqi Zhao, Rui Yao, and Zhiwen Shao. Incremental learning with neural networks for computer vision: a survey. *Artificial intelligence review*, 56(5):4557–4589, 2023.

Wenquan Lu, Jiaqi Zhang, Hugues Van Assel, and Randall Balestriero. Ditch the denoiser: Emergence of noise robustness in self-supervised learning from data curriculum. *arXiv preprint arXiv:2505.12191*, 2025.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022.

Guy Oren and Lior Wolf. In defense of the learning without forgetting for task incremental learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2209–2218, 2021.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024.

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

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6967–6977, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

Baoye Song, Shihao Zhao, Luyao Dang, Haoguang Wang, and Lin Xu. A survey on learning from data with label noise via deep neural networks. *Systems Science & Control Engineering*, 13(1):2488120, 2025.

Anirudh Subramanyam, Yuxin Chen, and Robert L Grossman. Scaling laws revisited: modeling the role of data quality in language model pretraining. *arXiv preprint arXiv:2510.03313*, 2025.

Hao Sun, Zhongyi Han, Hao Chen, Jindong Wang, Xin Gao, and Yilong Yin. From pretraining to pathology: How noise leads to catastrophic inheritance in medical models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yiyi Tao, Zhuoyue Wang, Hang Zhang, Lun Wang, and Jianxin Gu. Nevlp: Noise-robust framework for efficient vision-language pre-training. In *International Conference on Intelligent Computing*, pages 74–85. Springer, 2025.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022.

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

Santosh Vasa, Aditi Ramadwar, Jnana Rama Krishna Darabattula, Md Zafar Anwar, Stanislaw Antol, Andrei Vatavu, Thomas Monninger, and Sihao Ding. Autovdc: Automated vision data cleaning using vision-language models. *arXiv preprint arXiv:2507.12414*, 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

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

Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, and Mike Zheng Shou. Too large; data reduction for vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3147–3157, 2023.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024a.

Qiang Wang, Yuhang He, Songlin Dong, Xinyuan Gao, Shaokun Wang, and Yihong Gong. Non-exemplar domain incremental learning via cross-domain concept integration. In *European Conference on Computer Vision*, pages 144–162. Springer, 2024b.

Xiequn Wang, Zhan Zhuang, and Yu Zhang. Plan: Proactive low-rank allocation for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2909–2918, 2025a.

Yan Wang, Da-Wei Zhou, and Han-Jia Ye. Integrating task-specific and universal adapters for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 806–816, 2025b.

Zhixiang Wei, Guangting Wang, Xiaoxiao Ma, Ke Mei, Huaian Chen, Yi Jin, and Fengyun Rao. Hq-clip: Leveraging large vision-language models to create high-quality image-text datasets and clip models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22447–22456, 2025.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8363–8371, 2024.

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

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, Lijun Zhang, and De-Chuan Zhan. Dual consolidation for pre-trained model-based domain-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20547–20557, 2025b.

A Details of Experiments

A.1 Definitions of Common Evaluation Metrics

Let there be T tasks, and let $a_{i,j}$ denote the test accuracy on the j-th task after the model has completed learning the i-th task. The common metrics used in this paper are defined as follows:

- **Final Average Accuracy**: $A_T = \frac{1}{T}\sum_{j=1}^{T} a_{T,j}$.
- **Historical Average Accuracy**: $A = \frac{1}{T}\sum_{i=1}^{T}\left(\frac{1}{i}\sum_{j=1}^{i} a_{i,j}\right)$.
- **Forgetting Rate**: $F = \frac{1}{T-1}\sum_{j=1}^{T-1}\left(\max_{i\in\{j,\ldots,T\}} a_{i,j} - a_{T,j}\right)$.

A.2 Mechanistic Illustration of Inherited Output Contraction

Figure 9 provides an intuitive illustration of how pretraining noise induces an inherited output contraction. Under clean pretraining, semantically similar samples receive consistent supervision, encouraging well-separated class regions and a large logit gap between the true class and competing classes. This yields confident predictions and a stable margin structure for subsequent adaptation.

In contrast, noisy pretraining introduces conflicting supervision within local semantic regions. Nearby samples are driven toward different class directions, leading to overlap between class regions and reducing the true-class logit gap. This induces a low-margin regime where softmax outputs become flatter and ambiguous regions concentrate around cross-class boundaries. Since continual adaptation is initialized from pretrained parameters and operates with limited task-specific updates, these low-margin boundary structures are not fully reconfigured and are therefore carried into downstream adaptation, affecting both new-task learning and previous-task retention.

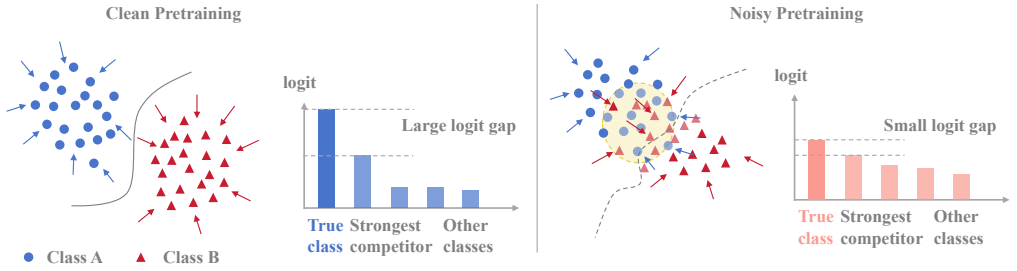


Figure 9: Mechanistic illustration of inherited output contraction. Clean pretraining induces well-separated class regions with large logit margins and stable decision boundaries. Noisy pretraining introduces conflicting supervision in local semantic regions, leading to a low-margin regime with compressed logit gaps and concentrated ambiguous boundaries. During continual adaptation, these boundary structures are inherited due to initialization from pretrained parameters and limited task-specific updates.

A.3 Detailed Results of Exploratory Experiments

Tables 4–7 provide task-wise accuracy used to compute the five metrics in Figure 2. Each row corresponds to a pretraining noise ratio.

Table 4: Detailed results of FFT on DomainNet.

	task1-Real World	task2-Clipart		task3-Painting			task4-Sketch			
	Real World	Real World	Clipart	Real World	Clipart	Painting	Real World	Clipart	Painting	Sketch
0%	81.73	66.49	71.88	59.31	47.47	67.82	50.71	48.84	43.86	65.19
5%	81.26	64.55	70.54	57.86	42.76	66.39	48.64	46.02	42.57	63.51
10%	81.20	64.29	70.25	56.69	43.26	66.20	47.99	46.26	42.03	63.18
20%	80.88	63.03	70.22	56.91	43.49	65.89	45.57	44.46	38.74	62.34
30%	80.06	61.32	69.00	54.86	40.81	65.18	44.65	42.89	38.20	60.79

Table 5: Detailed results of FFT on OfficeHome.

	task1-Real World	task2-Clipart		task3-Art			task4-Product			
	Real World	Real World	Clipart	Real World	Clipart	Art	Real World	Clipart	Art	Product
0%	81.77	70.64	71.48	70.87	57.85	70.16	72.82	51.32	59.47	89.98
5%	81.88	70.64	69.87	70.07	55.67	70.37	70.64	45.36	59.67	89.86
10%	80.96	69.27	68.38	66.06	51.89	67.08	70.18	42.15	57.41	88.96
20%	80.28	68.27	69.53	66.28	53.72	66.87	70.41	43.07	56.08	88.40
30%	80.16	73.51	68.38	63.76	48.34	63.58	64.79	37.23	52.88	88.74

Table 6: Detailed results of LoRA on DomainNet.

	task1-Real World	task2-Clipart		task3-Painting			task4-Sketch			
	Real World	Real World	Clipart	Real World	Clipart	Painting	Real World	Clipart	Painting	Sketch
0%	81.11	76.77	65.17	74.57	55.91	62.98	69.61	55.77	54.92	56.93
5%	80.47	75.98	64.35	73.24	54.68	61.70	67.71	53.53	52.87	53.77
10%	79.98	74.93	63.14	71.51	53.01	60.47	65.45	51.85	51.10	53.09
20%	79.05	73.77	62.69	71.47	51.19	59.73	65.98	50.37	49.92	52.79
30%	77.71	72.44	60.48	70.62	49.61	58.57	63.85	48.02	48.24	50.18

A.4 Detailed Results of Experiments with Asymmetric Label Noise

Tables 8 and 9 list the raw task-wise results for the method comparison under asymmetric synthetic label noise. These entries are the source values used to compute A_T, A, F, A_{TT}, and B in Figure 8.

A.5 Detailed Results of Experiments with Real-World Noise

Tables 10–13 report the raw task-wise accuracies for real-world pretrained model experiments. The five metrics summarized in Table 1 are all computed from these results, following the definitions in the main text.

For class-shift experiments, the original task-wise accuracy matrices are much larger because CUB-20 contains 10 sequential tasks. To ensure readability in the appendix, we report the average accuracy over all seen tasks at each adaptation stage. This compact representation preserves the learning trajectory across tasks and can be directly used to compute the metrics in Table 2. The results for CUB-20 and CUB-40 are shown in Tables 14 and 15, respectively.

Tables 16 and 17 provide the task-wise accuracies for experiments combining with existing continual adaptation methods, corresponding to the settings in Table 3.

B Additional Experiments

B.1 Different Orders

To examine whether the observed degradation is tied to a particular domain order, we additionally evaluate LoRA on DomainNet with two alternative task sequences. Figure 10 reports the order Real World $\rightarrow$ Painting $\rightarrow$ Clipart $\rightarrow$ Sketch, while Figure 11 reports the order Real World $\rightarrow$ Clipart $\rightarrow$ Sketch $\rightarrow$ Painting. Both figures show the same five metrics used in the main experiments.

Table 7: Detailed results of LoRA on OfficeHome.

	task1-Real World	task2-Clipart		task3-Art			task4-Product			
	Real World	Real World	Clipart	Real World	Clipart	Art	Real World	Clipart	Art	Product
0%	81.31	82.11	53.15	81.77	52.23	73.46	82.11	50.74	73.05	83.56
5%	80.16	80.50	49.94	80.39	47.65	67.70	80.73	48.11	68.11	81.53
10%	78.90	78.78	47.31	78.56	46.85	68.93	79.36	45.82	67.49	83.22
20%	78.10	76.49	47.31	77.41	48.11	66.26	78.44	46.16	65.64	80.41
30%	73.39	73.62	44.67	74.20	44.10	65.02	74.54	44.90	64.81	79.39

Table 8: Detailed noise-ratio results on DomainNet.

		task1-Real World	task2-Clipart		task3-Painting			task4-Sketch			
		Real World	Real World	Clipart	Real World	Clipart	Painting	Real World	Clipart	Painting	Sketch
0%	LoRA	81.11	76.77	65.17	74.57	55.91	62.98	69.61	55.77	54.92	56.93
	NML	82.46	77.04	70.28	75.16	60.65	67.34	69.05	60.26	56.87	63.37
	SKD	82.47	77.37	69.13	75.32	60.63	64.24	69.87	60.59	56.43	59.28
	Ours	82.39	79.34	69.95	79.14	63.44	66.78	75.91	62.61	59.93	63.33
5%	LoRA	80.47	75.98	64.35	73.24	54.68	61.70	67.71	53.53	52.87	53.77
	NML	81.62	76.64	67.98	73.94	57.52	65.41	67.45	57.19	54.85	60.32
	SKD	81.64	76.98	66.76	73.89	57.03	61.91	68.10	57.36	54.17	55.14
	Ours	81.80	78.59	68.52	77.91	60.37	64.44	74.27	61.09	57.74	60.45
10%	LoRA	79.98	74.93	63.14	71.51	53.01	60.47	65.45	51.85	51.10	53.09
	NML	81.11	76.19	66.51	73.63	56.59	64.32	67.23	55.59	54.18	58.34
	SKD	81.12	76.30	65.16	73.68	55.85	60.87	67.40	55.39	53.08	52.61
	Ours	81.25	78.23	67.04	77.71	59.44	63.68	74.71	59.90	56.59	58.55
20%	LoRA	79.05	73.77	62.69	71.47	51.19	59.73	65.98	50.37	49.92	52.79
	NML	80.33	74.99	65.76	72.40	54.49	63.30	65.64	54.40	52.52	58.03
	SKD	80.38	75.20	64.39	72.53	54.52	60.26	66.29	54.42	52.30	52.54
	Ours	80.92	77.73	67.32	76.94	60.28	63.78	72.72	59.80	55.79	58.74
30%	LoRA	77.71	72.44	60.48	70.62	49.61	58.57	63.85	48.02	48.24	50.18
	NML	79.03	74.50	63.39	71.56	52.41	61.47	64.55	51.71	50.19	55.87
	SKD	79.05	74.63	61.80	71.43	51.70	57.71	64.98	51.73	49.50	50.06
	Ours	79.72	76.30	66.18	75.97	57.96	62.09	71.51	58.29	54.75	56.87

Table 9: Detailed noise-ratio results on OfficeHome.

		task1-Real World	task2-Clipart		task3-Art			task4-Product			
		Real World	Real World	Clipart	Real World	Clipart	Art	Real World	Clipart	Art	Product
0%	LoRA	81.31	82.11	53.15	81.77	52.23	73.46	82.11	50.74	73.05	83.56
	NML	85.32	83.94	67.01	82.80	64.15	76.95	83.49	60.37	74.69	90.32
	SKD	84.40	85.21	61.97	84.52	58.19	76.34	86.24	58.53	74.49	87.73
	Ours	83.94	84.63	69.07	85.32	66.21	78.81	85.78	62.89	77.16	90.65
5%	LoRA	80.16	80.50	49.94	80.39	47.65	67.70	80.73	48.11	68.11	81.53
	NML	84.86	83.03	64.38	81.65	61.40	75.51	81.65	57.50	72.63	89.30
	SKD	83.37	83.83	59.68	83.03	57.39	70.99	84.06	57.85	69.34	86.60
	Ours	84.40	83.26	67.58	83.37	65.75	74.69	83.72	59.79	75.10	89.86
10%	LoRA	78.90	78.78	47.31	78.56	46.85	68.93	79.36	45.82	67.49	83.22
	NML	84.63	83.94	61.63	81.54	60.02	73.66	82.11	56.36	71.81	88.85
	SKD	83.83	82.34	57.04	81.77	53.26	71.40	82.80	53.15	71.19	84.46
	Ours	83.37	82.57	63.57	82.80	63.12	76.13	83.37	56.93	71.81	87.39
20%	LoRA	78.10	76.49	47.31	77.41	48.11	66.26	78.44	46.16	65.64	80.41
	NML	82.91	80.05	62.31	78.90	58.76	71.19	80.16	54.18	67.70	87.95
	SKD	81.88	80.85	58.08	80.39	54.41	67.08	81.65	54.30	64.61	83.67
	Ours	82.68	81.54	64.60	81.54	61.28	72.43	82.68	56.13	70.99	87.16
30%	LoRA	73.39	73.62	44.67	74.20	44.10	65.02	74.54	44.90	64.81	79.39
	NML	80.73	79.13	59.79	77.06	58.99	69.34	78.67	50.97	66.05	87.27
	SKD	80.85	79.70	55.44	78.78	52.69	66.46	79.13	53.04	66.05	84.35
	Ours	83.37	81.42	61.74	80.96	59.11	72.22	82.11	55.44	71.81	86.26

Table 10: Detailed task-wise results on DomainNet.

	task1-Real World	task2-Clipart		task3-Painting			task4-Sketch			
	Real World	Real World	Clipart	Real World	Clipart	Painting	Real World	Clipart	Painting	Sketch
LoRA	91.54	89.62	83.57	88.85	80.57	80.57	88.14	81.01	76.17	78.40
NML	91.77	94.02	84.50	89.05	81.49	81.98	86.94	82.31	75.73	79.45
SKD	91.68	90.08	83.87	89.01	81.28	79.51	87.26	82.24	75.49	76.48
Ours	91.84	90.41	84.99	89.79	82.06	81.96	89.22	82.18	78.57	79.22
Ours (Feature Projection)	91.75	90.18	84.89	88.77	81.55	82.15	87.31	81.58	75.69	80.02

Table 11: Detailed task-wise results on OfficeHome.

	task1-Real World	task2-Clipart		task3-Art			task4-Product			
	Real World	Real World	Clipart	Real World	Clipart	Art	Real World	Clipart	Art	Product
LoRA	91.63	92.78	77.55	93.00	78.47	82.92	93.58	78.47	84.98	95.16
NML	94.04	92.78	81.67	92.55	79.95	87.65	92.32	79.38	86.01	95.38
SKD	93.23	92.55	78.12	92.89	76.17	87.24	92.78	78.69	86.42	95.38
Ours	93.92	94.04	86.03	94.84	84.19	90.53	95.07	83.96	88.89	95.95
Ours (Feature Projection)	94.72	91.97	84.42	93.92	82.59	89.71	94.15	81.67	87.45	96.40

Table 12: Detailed task-wise results on CDDB.

	task1-san	task2-whichfaceisreal		task3-biggan			task4-wild				task5-gaugan				
	san	san	whichfaceisreal	san	whichfaceisreal	biggan	san	whichfaceisreal	biggan	wild	san	whichfaceisreal	biggan	wild	gaugan
LoRA	60.00	53.33	69.25	50.00	67.25	70.00	47.78	71.25	64.12	52.01	48.89	50.00	84.25	49.83	95.40
NML	64.44	56.67	70.00	56.67	65.00	69.62	50.00	51.10	52.25	53.32	52.22	51.25	76.38	48.52	90.75
SKD	64.44	63.33	57.75	62.22	54.50	57.25	51.11	50.25	50.25	51.82	62.22	57.00	60.75	52.54	62.80
Ours	67.78	52.20	67.25	54.44	63.75	78.12	65.56	72.75	79.25	50.36	58.89	50.25	95.00	49.64	98.75
Ours (Feature Projection)	66.67	61.11	78.50	57.78	61.00	74.75	50.00	53.50	60.50	53.56	52.22	52.00	91.62	49.20	97.60

Table 13: Detailed task-wise results on TerraIncognita.

	task1-Location100	task2-Location38		task3-Location43			task4-Location46			
	Location100	Location100	Location38	Location100	Location38	Location43	Location100	Location38	Location43	Location46
LoRA	80.26	70.08	76.42	72.13	71.41	54.73	70.77	67.83	55.72	53.70
NML	86.61	71.31	75.37	59.22	63.08	58.87	70.01	53.29	45.27	61.66
SKD	86.07	77.32	57.39	77.05	58.34	51.16	73.84	50.39	46.27	54.74
Ours	84.02	70.01	80.35	71.52	74.58	63.76	68.37	64.89	65.26	63.07
Ours (Feature Projection)	89.89	75.41	83.79	57.86	76.83	71.14	67.14	70.04	66.00	73.86

Table 14: Task-stage average accuracy on CUB-20. Each A_i denotes the average accuracy over all seen tasks after learning task i.

Method	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
LoRA	85.24	74.93	63.11	59.19	60.64	57.45	52.90	50.46	46.09	47.08
NML	93.79	84.46	75.52	73.58	74.73	72.64	69.90	68.80	62.68	66.22
SKD	93.79	82.13	77.20	74.30	73.03	68.54	60.96	60.13	54.49	54.88
Ours	90.87	83.39	76.66	75.45	75.60	74.41	70.80	71.04	67.50	67.19
Ours (Feature Projection)	93.98	85.61	77.98	76.77	77.81	76.23	72.06	71.56	65.92	68.82

Table 15: Task-stage average accuracy on CUB-40. Each A_i denotes the average accuracy over all seen tasks after learning task i.

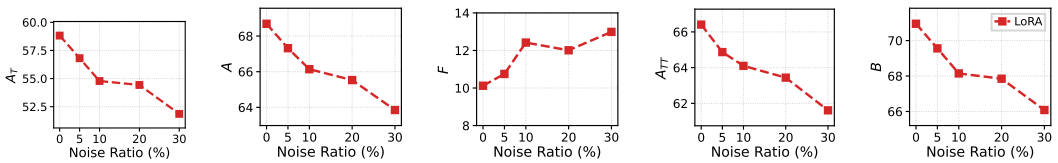
Method	A_1	A_2	A_3	A_4	A_5
LoRA	77.16	68.22	66.68	61.53	57.22
NML	86.06	76.02	76.68	73.44	72.74
SKD	85.78	78.03	75.83	71.18	67.93
Ours	86.15	81.77	81.27	78.37	76.39
Ours (Feature Projection)	86.88	80.29	80.18	76.97	76.06

Table 16: Detailed task-wise combination results on DomainNet.

	task1-Real World	task2-Clipart		task3-Painting			task4-Sketch			
	Real World	Real World	Clipart	Real World	Clipart	Painting	Real World	Clipart	Painting	Sketch
APER-Adapter	86.34	86.36	79.12	86.11	79.06	74.16	86.06	78.77	74.05	70.00
APER-Adapter + Ours	86.38	86.41	79.25	86.20	79.20	74.05	86.16	78.90	73.92	69.90
TUNA	89.20	87.26	80.68	86.24	80.14	76.22	85.87	79.90	75.25	72.39
TUNA + Ours	90.21	88.62	83.18	87.55	81.82	78.02	86.95	81.34	76.33	73.88

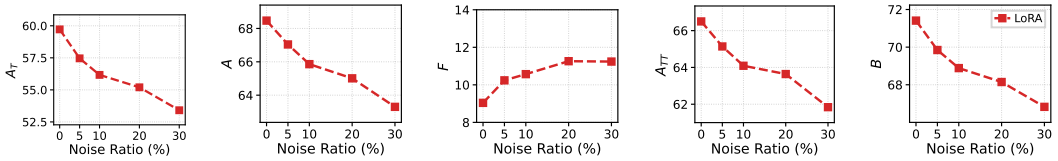
Table 17: Detailed task-wise combination results on OfficeHome.

	task1-Real World	task2-Clipart		task3-Art			task4-Product			
	Real World	Real World	Clipart	Real World	Clipart	Art	Real World	Clipart	Art	Product
APER-Adapter	92.09	92.09	76.29	91.86	76.40	82.51	91.28	76.40	82.72	93.69
APER-Adapter + Ours	92.66	92.66	78.12	92.66	78.35	83.95	91.97	78.35	83.74	93.81
TUNA	88.76	92.32	77.43	91.74	78.24	81.89	91.40	78.47	82.72	93.24
TUNA + Ours	93.00	92.78	80.87	93.23	80.99	84.57	92.43	81.33	84.98	94.82



(a) $A_T \uparrow$ (b) $A \uparrow$ (c) $F \downarrow$ (d) $A_{TT} \uparrow$ (e) $B \uparrow$

Figure 10: LoRA results on DomainNet under the task order Real World → Painting → Clipart → Sketch across five metrics.

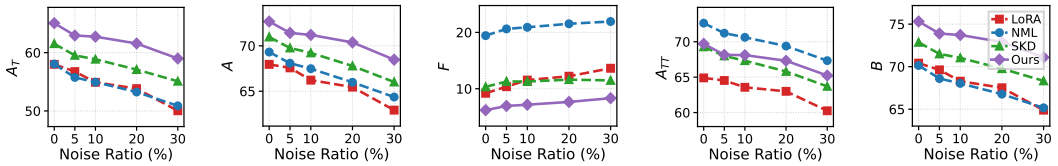


(a) $A_T \uparrow$ (b) $A \uparrow$ (c) $F \downarrow$ (d) $A_{TT} \uparrow$ (e) $B \uparrow$

Figure 11: LoRA results on DomainNet under the task order Real World $\rightarrow$ Clipart $\rightarrow$ Sketch $\rightarrow$ Painting across five metrics.

B.2 Random Noise

In addition to asymmetric label noise, we further pretrain another set of ViT-B/16 models with randomly corrupted (symmetric) label noise and evaluate multiple methods on DomainNet. Figure 12 reports the corresponding five metrics, providing a direct comparison with the asymmetric-noise setting in Figure 2. The results show that symmetric noise exhibits a similar impact pattern to asymmetric noise, while our method consistently remains effective under this setting.



(a) $A_T \uparrow$ (b) $A \uparrow$ (c) $F \downarrow$ (d) $A_{TT} \uparrow$ (e) $B \uparrow$

Figure 12: Results under symmetric label noise on DomainNet across five metrics.

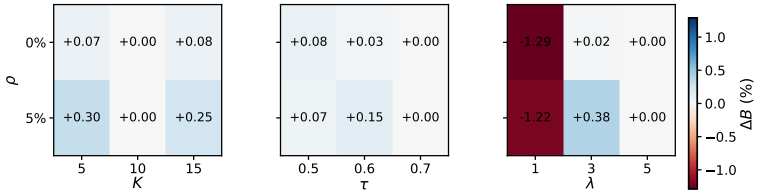


Figure 13: Hyperparameter sensitivity analysis on DomainNet. The heatmap reports the change in B under different values of λ, τ, and K.

B.3 Hyperparameter Analysis

We further analyze the sensitivity of the proposed method to its main hyperparameters. Figure 13 reports the change in the comprehensive metric B on DomainNet when varying λ, τ, and K around the default setting.

C Limitations

This paper focuses on representative label noise and widely used continual adaptation benchmarks. The current study provides limited discussion of higher noise levels, including the extreme case of 100% noise, and the behavior of continual adaptation under such severe corruption remains to be further investigated. Although experiments cover both synthetic and real-world noisy pretrained models, the current study is limited to continual visual classification, and further validation on other tasks such as detection, segmentation, and language tasks remains future work.

D Broader Impacts

This work studies how pretraining noise affects foundation models during downstream continual adaptation and proposes ABC to improve robustness under such conditions. The method may benefit applications where models are repeatedly adapted to changing environments, such as medical diagnosis, autonomous driving, and long-term deployed recognition systems. However, emphasizing low-confidence ambiguous regions may also affect rare or minority patterns if they are incorrectly treated as noise-induced uncertainty. In high-stakes applications, we recommend combining the proposed method with fairness assessments, uncertainty monitoring, and human-in-the-loop validation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the appendix (Section C).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All necessary experimental details are provided in Section 3.1, Section 5, and the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the code in the supplementary material, and plan to release them publicly after final organizational preparations are completed.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are provided in Section 3.1, Section 5, and the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experimental results are reported under consistent settings, with all experiments repeated three times and results averaged across runs. Detailed information is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The required computational resource details are provided in Section 3.1 and Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussed in Section D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve high-risk model or dataset release requiring specific safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used assets are properly credited and their licenses and terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed descriptions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM not used in the core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.