CAR: CONTEXTUAL ADAPTER ROUTING FOR REHEARSAL-FREE CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern rehearsal-free class-incremental learning (CIL) increasingly builds on pretrained models augmented with task-specialized prompts, adapters, or other modular components. While substantial effort has focused on how such modules are learned and stored, inference is still typically performed independently for each test sample, discarding contextual evidence that may be available across nearby observations. We introduce **Contextual Adapter Routing (CAR)**, a frozen context-conditioned inference strategy that exploits a small unlabeled coherent test context without replay, task labels, retraining, or test-time parameter updates. CAR observes that task-specialized modules contain weak but useful task-discriminative information in their predictive uncertainty: although this signal is noisy for individual samples, aggregating it across a coherent context yields substantially more reliable module preferences. CAR converts this contextual preference into a soft task prior while retaining image-specific all-class prediction, thereby improving routing without replacing the underlying continual learner. Across multiple CIL benchmarks and modular learners, CAR improves average continual and final-stage accuracy by up to 11.14 and 13.49 percentage points, respectively. Controlled analyses show that the gain depends on contextual coherence: distinct coherent observations provide substantially stronger evidence than same-image augmentations or unsupervised grouping of shuffled samples. We further study unreliable contexts and use routing confidence to fall back toward native sample-wise inference when contextual evidence is weak. Theoretically, when observations share a local shifted distribution and the appropriate module has a positive expected entropy margin, the probability of routing error decreases exponentially with context size. Together, these results establish *frozen context-conditioned inference* as a complementary design axis for modular continual learning: preserve native sample-wise inference when reliable context is unavailable, while exploiting coherent unlabeled context when it is informative.

1 INTRODUCTION

In class-incremental learning (CIL), stages introduce new classes while the final model must classify over all classes seen so far, without task identity at test time (Rebuffi et al., 2017). Modern rehearsal-free methods increasingly build on pretrained models and store continual knowledge in lightweight prompts, adapters (Houlsby et al., 2019), low-rank updates (Hu et al., 2022), classifiers, or task-specific subspaces (Zhou et al., 2025; McDonnell et al., 2023; Wang et al., 2022b;a; Smith et al., 2023; Liang & Li, 2024; Zhou et al., 2024; He et al., 2025). These strategies substantially reduce interference during continual training, but modular methods introduce a second, comparatively underexplored question: *how should a frozen bank of specialized components be used when the task identity of an incoming sample is unknown?*

Standard CIL evaluation normally treats each test observation independently. Test-time adaptation (TTA), by contrast, exploits unlabeled test data by updating model parameters, normalization statistics, prompts, prototypes, or other internal state (Wang et al., 2021). We study the intermediate setting illustrated in Fig. 1: **can a continual learner exploit nearby unlabeled observations while keeping the trained model fixed?** This defines a distinct inference regime along two axes. Standard CIL keeps the learner frozen but does not allow neighboring observations to influence one another; TTA exploits test context but adapts parameters or state; our setting exploits test context while leav-

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

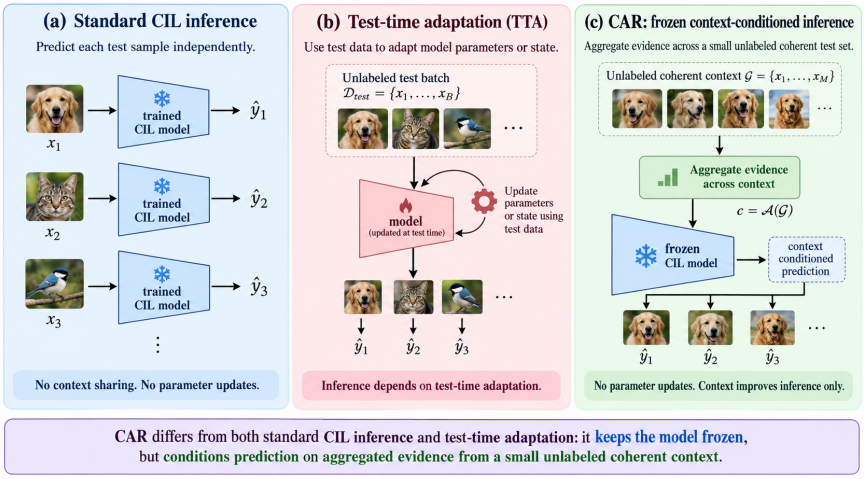


Figure 1: **Inference settings.** Standard CIL predicts each sample independently. Test-time adaptation (TTA) exploits unlabeled test data by updating parameters or model state. CAR instead performs *frozen context-conditioned inference*: a small unlabeled coherent context influences prediction while all learned parameters remain fixed.

ing the trained continual learner unchanged. This separation is useful because improvements can be attributed to inference-time evidence aggregation rather than additional optimization. Appendix B provides a detailed comparison of standard sample-wise CIL, test-time adaptation, and the frozen context-conditioned setting studied here.

Such a setting is practically relevant because many data streams possess weak local structure. Examples include adjacent video frames, acquisition bursts, neighboring microscopy or remote-sensing crops, and short operating regimes of a physical system. These observations need not share labels, nor do they reveal task identity. Instead, they provide a small unlabeled context around the current observation that may contain information about the underlying local data-generating regime. We therefore ask whether context itself can be treated as an inference-time resource, analogous to how replay memory, parameter isolation, or modular expansion are treated as training-time resources in continual learning.

Our focus is the broad family of *modular rehearsal-free CIL* methods that retain task-specialized components. Representative examples include adapter-based approaches such as EASE (Zhou et al., 2024) and CL-LoRA (He et al., 2025), which differ substantially in how specialized representations are learned and organized but share an inference-time property important for our purpose: multiple frozen task-dependent components can be evaluated on an incoming observation. This is the interface required by our approach. Rather than changing how those modules are trained, we ask whether their existing predictive responses can be used more effectively at inference. We instantiate the same inference mechanism on both CL-LoRA and EASE to test whether the resulting behavior transfers across different modular CIL designs.

We introduce **Contextual Adapter Routing (CAR)** to realize this idea. CAR emerged from a broader learned inference-controller study in which a controller was trained to choose among adapter-dependent inference actions. Across experiments, the learned policy repeatedly collapsed toward selecting the adapter with the lowest task-local predictive entropy. This observation suggested that frozen task-specific modules, although never trained explicitly as task identifiers, already expose weak task-discriminative compatibility signals through their predictive responses. At the level of a single image, however, these signals are noisy and routing remains unreliable. Appendix E documents the controller formulation and diagnostic results that led to this observation.

CAR makes this weak signal useful by aggregating it across a small context $G = x_1, \ldots, x_M$. Each specialized module is evaluated on the same context, its task-local predictive uncertainty is aggregated, and the resulting contextual preference is converted into a soft task prior. Importantly, this prior does not replace the image-specific classifier with a hard task-ID decision: every observation is still classified individually over all classes seen so far. CAR therefore improves module selection

while preserving all-class prediction. The entire procedure is frozen and requires no replay, task labels, retraining, gradient updates, or test-time parameter modification.

CAR is also *opportunistic rather than mandatory*. When deployment provides no trustworthy grouping of observations, the underlying continual learner can retain its original sample-wise inference unchanged. When an externally available local context exists—for example, a temporal burst, spatial neighborhood, or acquisition episode—the same frozen model can use CAR to aggregate additional evidence and improve routing reliability. We further consider unreliable contexts by using the contextual routing margin as a confidence signal: when contextual evidence becomes weak, confidence-gated CAR can fall back toward native inference. This extension is intended as a practical robustness mechanism rather than a formal no-harm guarantee.

The gains are substantial despite the absence of adaptation. With only eight coherent observations, CAR improves average continual accuracy by up to **11.14 percentage points** and final-stage accuracy by up to **13.49 percentage points**. On ImageNet-R, final-stage accuracy increases from 69.48% to 82.97%, while on CIFAR-100 it increases from 87.94% to 98.06%. The same context-size trend also appears when CAR is applied to EASE, indicating that the effect is not specific to CL-LoRA or to a single adapter architecture. Controlled analyses further show that the gain does not arise merely from processing multiple inputs: repeated augmentations of the same image and unsupervised grouping of shuffled observations provide much weaker routing signals than distinct coherent observations.

Our contributions are fourfold: (i) we formulate *frozen context-conditioned inference* as a distinct CIL inference setting that lies between independent frozen inference and parameter- or state-updating TTA; (ii) we introduce CAR, a training-free contextual routing mechanism applicable to modular continual learners through their frozen predictive responses, and instantiate it on both CL-LoRA and EASE; (iii) we demonstrate consistent gains across multiple CIL benchmarks and provide controlled diagnostics showing when coherent multi-sample context is informative and when it is not; and (iv) we show theoretically that, when observations share a local distribution and the appropriate module has a positive expected entropy margin, the probability of routing error decreases exponentially with context size.

2 Related Work

Rehearsal-free CIL with pretrained models. Recent class-incremental learning (CIL) increasingly relies on large pretrained backbones and parameter-efficient adaptation (Houlsby et al., 2019; Hu et al., 2022) rather than repeatedly updating the full network or storing exemplars from previous tasks. One line of work uses a largely shared or non-expanding representation. SimpleCIL shows that a frozen pretrained encoder combined with simple prototype-style classification can already be highly competitive, while APER strengthens this paradigm through adaptation of the pretrained representation (Zhou et al., 2025). RanPAC instead constructs a high-dimensional random-projection classifier on top of pretrained features to improve class separation while avoiding rehearsal (McDonnell et al., 2023). InfLoRA introduces low-rank adaptation for continual learning while constraining updates to reduce interference with previously acquired knowledge (Liang & Li, 2024). More recent LoRA-based methods explicitly target scalability: SD-LoRA continually decouples the magnitude and direction of low-rank updates and ultimately uses a single final model without module selection at inference (Wu et al., 2025), while EWC-LoRA regularizes sequential LoRA adaptation to preserve previously important parameters (Zheng et al., 2026). MoAL combines an analytic classifier with momentum-based adapter interpolation to balance knowledge retention and adaptation to new classes (Gao et al., 2025), and VILA similarly seeks parameter-efficient continual adaptation with a compact shared model (Zhao et al., 2026). These methods primarily ask how a pretrained representation can be updated, consolidated, or reused across a continual sequence.

A complementary family retains explicit task- or component-specific parameters. L2P maintains a learned prompt pool and retrieves input-conditioned prompts without requiring task identity at test time (Wang et al., 2022b). DualPrompt separates prompts into task-invariant and task-specific components (Wang et al., 2022a), while CODA-Prompt learns decomposed prompt components that are combined with input-conditioned attention (Smith et al., 2023). EASE learns a new lightweight adapter for each task, creating task-specific feature subspaces and later combining information across the expanding ensemble (Zhou et al., 2024). CL-LoRA similarly maintains both task-shared

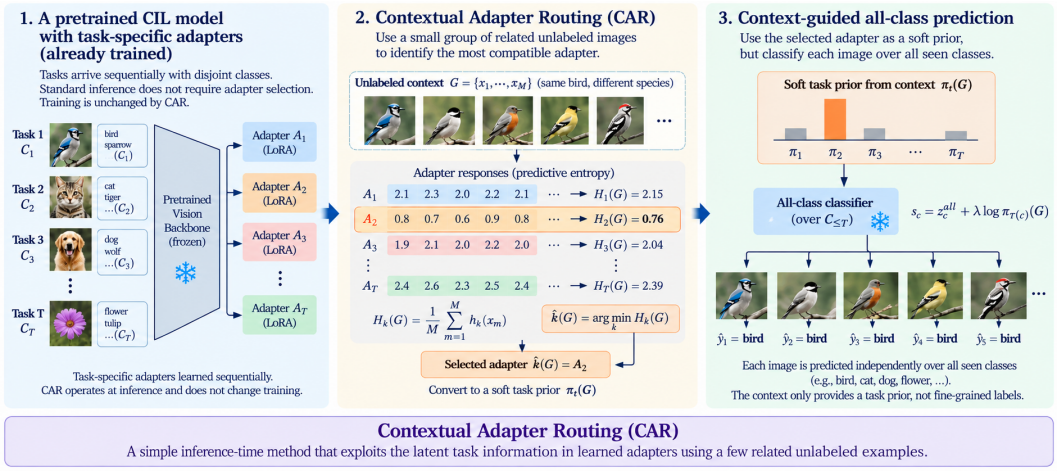


Figure 2: **Overview of Contextual Adapter Routing (CAR).** CAR leaves continual training unchanged and modifies only inference: a small unlabeled context is evaluated by the frozen adapter bank, per-sample compatibility evidence is aggregated to select the most compatible adapter, and prediction proceeds without task IDs or model updates.

and task-specific low-rank adapters to capture transferable and task-dependent information (He et al., 2025). SEMA, DIA, and ACMap explore related modular designs for retaining, merging, or consolidating specialized components as the task sequence grows (Wang et al., 2025a; Li et al., 2025; Fukuda et al., 2025). The common emphasis across these approaches is therefore on *how knowledge is stored and learned* during continual training. CAR addresses an orthogonal question: given an already-trained modular continual learner, how should its specialized components be used at test time? CAR leaves the backbone, adapters, prompts, and classifiers unchanged and modifies only frozen inference.

Routing, uncertainty, and test-time inference. Once task-specialized modules are maintained, prediction additionally requires deciding which module, or combination of modules, should influence a test example. The closest connection to CAR is TUNA, which performs task-adapter selection independently for each test sample using predictive entropy as a task-uncertainty signal (Wang et al., 2025b). This is closely related to the broader observation that predictive confidence and entropy contain useful information about whether a model is operating on familiar data, although neural-network confidence is not necessarily calibrated (Hendrycks & Gimpel, 2017; Guo et al., 2017). CAR starts from the same general intuition but makes a different inference assumption: task evidence need not be estimated from each image in isolation. Instead, when a small set of unlabeled observations is believed to arise from a common local context, CAR aggregates module-level uncertainty across those observations to obtain a more stable contextual routing signal. The resulting adapter preference is not used as a hard task-ID oracle; it is converted into a soft prior that is combined with each image's native all-class prediction.

This distinction also separates CAR from test-time adaptation (TTA). TTA methods use unlabeled test observations to compensate for distribution shift by modifying the model during deployment. For example, Tent adapts normalization parameters through test-time entropy minimization (Wang et al., 2021); CONTRAST uses online self-training and consistency mechanisms for continual adaptation under changing test distributions (Ahmed et al., 2024); and recent approaches such as FreeTTA and ReTTA similarly modify parameters, representations, statistics, or internal state using test data (Dai & Yang, 2025; Park et al., 2025). Even training-free test-time methods can alter features, prototypes, or inference state based on the incoming stream (Wang et al., 2025c). CAR has a different objective and mechanism: it performs *frozen module selection* for a continual learner. No parameter, normalization statistic, prototype, or learned state is updated; the context is used only to estimate which already-learned module is most compatible with the current local observations. Thus, standard CIL predicts samples independently with the learned continual model, TTA adapts the model using test data, whereas CAR performs frozen context-conditioned inference over an existing modular model. Appendix B summarizes these distinctions.

3 PROBLEM SETTING

We consider rehearsal-free class-incremental learning over a sequence of T tasks, $\{\mathcal{D}_t\}_{t=1}^{T}$. At stage t, the learner receives

$$\mathcal{D}_t = \{(x_n^{(t)}, y_n^{(t)})\}_{n=1}^{N_t}, \qquad y_n^{(t)} \in \mathcal{C}_t,$$

where $\mathcal{C}_t$ is the class set introduced at task t. We assume the standard CIL setting with mutually disjoint class sets, $\mathcal{C}_i \cap \mathcal{C}_j = \varnothing$ for $i \neq j$. After stage t, the complete set of observed classes is

$$\mathcal{C}_{\leq t} = \bigcup_{k=1}^{t} \mathcal{C}_k. \tag{1}$$

Previously observed raw training samples are not replayed. At test time, the task identity is unavailable; hence prediction must be performed over the entire seen-class space $\mathcal{C}_{\leq t}$.

Modular continual learner. We focus on pretrained continual learners that retain task-specialized modules. After stage t, the model consists of a shared backbone f_θ and a bank of $K = t$ task-specific adapters,

$$\mathcal{A}_{\leq t} = \{A_k\}_{k=1}^{K} = \{A_1, \ldots, A_K\},$$

where adapter A_k is associated with class block $\mathcal{C}_k$.

For an input x, let $z_{k,c}(x)$ denote the logit for class c when candidate adapter A_k is active. Separately, let $z_c^{\mathrm{all}}(x)$ denote the native all-class logit of the underlying continual learner for $c \in \mathcal{C}_{\leq t}$.

Under standard sample-wise CIL inference,

$$\hat{y}^{\mathrm{native}}(x) = \arg\max_{c \in \mathcal{C}_{\leq t}} z_c^{\mathrm{all}}(x). \tag{2}$$

Context-conditioned inference. Instead of treating every test sample independently, we assume access to a small unlabeled context

$$G = \{x_i\}_{i=1}^{M} = \{x_1, \ldots, x_M\},$$

where M is the context size. The membership of G is determined without test labels, e.g., through temporal proximity, spatial adjacency, or acquisition-level grouping.

CAR uses G only to infer a shared routing signal. Each observation x_i still receives its own prediction over the full label space $\mathcal{C}_{\leq t}$. All backbone, adapter, and classifier parameters remain fixed during inference.

4 CONTEXTUAL ADAPTER ROUTING

CAR consists of three steps: (i) compute task-local adapter evidence for each observation, (ii) aggregate this evidence over the context, and (iii) convert the resulting routing decision into a soft prior for native all-class prediction.

4.1 FROM LEARNED INFERENCE CONTROL TO CAR

CAR was motivated by a broader learned inference-controller study. Given a frozen bank of task-specialized adapters, the controller could choose among adapter-dependent inference actions, yet its learned policy repeatedly concentrated on selecting the adapter with minimum task-local predictive entropy. This behavior suggested that frozen modules already expose a useful compatibility signal, but one that is too noisy for reliable single-image routing. CAR removes the learned controller and instead stabilizes this signal by aggregating it across a coherent context, introducing no trainable inference component or test-time update. Appendix E gives the complete controller formulation and diagnostics; the final CAR method does not use the controller.

4.2 TASK-LOCAL ADAPTER EVIDENCE

For observation x_i and candidate adapter A_k, we form a predictive distribution only over the adapter's associated class block:

$$p_{k,c}(x_i) = \frac{\exp(z_{k,c}(x_i))}{\sum_{d \in \mathcal{C}_k} \exp(z_{k,d}(x_i))}, \qquad c \in \mathcal{C}_k, \tag{3}$$

In all reported experiments, entropy is computed directly from the task-local logits with no additional routing temperature ($T = 1$).

We quantify adapter uncertainty using predictive entropy,

$$h_k(x_i) = -\sum_{c \in \mathcal{C}_k} p_{k,c}(x_i) \log p_{k,c}(x_i). \tag{4}$$

A smaller $h_k(x_i)$ indicates that adapter A_k produces a more concentrated prediction within its own class block. Importantly, $p_{k,c}(x_i)$ is not interpreted as a posterior over tasks; $h_k(x_i)$ is used only as an adapter–input compatibility score. Predictive uncertainty can become miscalibrated under distribution shift (Ovadia et al., 2019); CAR therefore uses entropy only as relative compatibility evidence rather than as a calibrated task posterior.

All reported CAR experiments use the raw entropy in Eq. (4); we do not normalize by task-block cardinality. The benchmarks considered here use fixed class increments within an evaluation, so the candidate task-local heads have equal cardinality at a given stage.

4.3 CONTEXTUAL AGGREGATION AND ROUTING

Given context G, CAR averages the entropy of each candidate adapter across the M observations:

$$H_k(G) = \frac{1}{M} \sum_{i=1}^{M} h_k(x_i). \tag{5}$$

The context-selected adapter is

$$\hat{k}(G) = \operatorname*{arg\,min}_{k \in \{1,\dots,K\}} H_k(G). \tag{6}$$

When $M = 1$, CAR reduces exactly to sample-wise minimum-entropy routing.

For analysis, let $H_{(1)}(G)$ and $H_{(2)}(G)$ denote the smallest and second-smallest contextual entropy scores. We define the routing margin as

$$\Delta(G) = H_{(2)}(G) - H_{(1)}(G). \tag{7}$$

4.4 SOFT CONTEXTUAL PRIOR AND ALL-CLASS PREDICTION

A hard routing decision is insufficient for CIL because all previously seen classes must remain valid candidates. We therefore convert $\hat{k}(G)$ into a smoothed task prior:

$$\pi_k(G) = \begin{cases} 1 - \epsilon, & k = \hat{k}(G), \\ \dfrac{\epsilon}{K - 1}, & k \neq \hat{k}(G), \end{cases} \tag{8}$$

for $K > 1$, with $\pi_1(G) = 1$ when $K = 1$.

Let $\tau(c)$ denote the unique task index satisfying $c \in \mathcal{C}_{\tau(c)}$. CAR modifies the native all-class score as

$$s_c(x_i \mid G) = z_c^{\text{all}}(x_i) + \lambda \log \pi_{\tau(c)}(G), \tag{9}$$

where $\lambda \geq 0$ controls the strength of contextual evidence. We use $\lambda = 1$ in all reported experiments. The prior biases the native all-class logits; CAR does not replace the full classifier with a task-local head.

The final prediction is

$$\hat{y}_i = \operatorname*{arg\,max}_{c \in \mathcal{C}_{\leq t}} s_c(x_i \mid G). \tag{10}$$

4.5 CONFIDENCE-GATED FALLBACK

When context quality is uncertain, we reuse the routing margin $\Delta(G)$ from Eq. (7) as a confidence signal and activate CAR only when $\Delta(G) \geq \tau$:

$$\hat{y}_i^{\text{gate}} = \begin{cases} \hat{y}_i^{\text{CAR}}, & \Delta(G) \geq \tau, \\ \hat{y}_i^{\text{native}}, & \Delta(G) < \tau. \end{cases} \tag{11}$$

This optional gate introduces no learned parameters and provides a fallback toward native sample-wise inference when contextual evidence is ambiguous; we make no formal no-harm guarantee. Its confidence-based fallback is related in spirit to selective prediction (Geifman & El-Yaniv, 2019), but here confidence chooses between two frozen inference modes rather than rejecting a sample.

4.6 ALGORITHMIC SUMMARY

CAR computes the KM task-local adapter–observation responses, averages entropy within the context, forms the contextual prior, and predicts each image over all seen classes. The implementation collects task-local responses and native all-class logits in shared batched forwards, requiring no selected-adapter second pass, backward pass, or parameter update. Pseudocode appears in Appendix A.

4.7 RELIABLE ROUTING UNDER COHERENT DISTRIBUTION SHIFT

A Hoeffding-style concentration argument (Hoeffding, 1963) formalizes how contextual averaging stabilizes routing. Let Q denote the local test distribution generating a contextual episode $G = \{x_1, \ldots, x_M\}$ with $x_i \overset{\text{i.i.d.}}{\sim} Q$. The distribution Q may differ from the training distribution. Let $k^\star$ denote the adapter most compatible with Q, and for each competitor $j \neq k^\star$ define

$$D_j(X) = h_j(X) - h_{k^\star}(X), \qquad X \sim Q. \tag{12}$$

Theorem 1 (Reliable contextual routing under a local entropy margin)**.** *Assume that, for every $j \neq k^\star$, the entropy-gap random variable $D_j(X)$ is bounded in an interval of width R_j and satisfies*

$$\mathbb{E}_{X \sim Q}[D_j(X)] \geq \gamma_j > 0. \tag{13}$$

Then CAR selects the correct adapter with probability at least

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - \sum_{j \neq k^\star} \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right). \tag{14}$$

Corollary 1 (Context size for reliable routing)**.** *Let*

$$\gamma = \min_{j \neq k^\star} \gamma_j, \qquad R = \max_{j \neq k^\star} R_j.$$

Then

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - (K-1)\exp\left(-\frac{2M\gamma^2}{R^2}\right). \tag{15}$$

Therefore, for any $\delta \in (0, 1)$, it is sufficient to choose

$$M \geq \frac{R^2}{2\gamma^2} \log\left(\frac{K-1}{\delta}\right) \tag{16}$$

to guarantee

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - \delta. \tag{17}$$

The bound formalizes coherent aggregation: routing error decreases exponentially in M under a positive local entropy margin, while the sufficient context size grows only logarithmically with the number of competing adapters. The guarantee is local to Q: it permits Q to differ from training data, but does not guarantee robustness to arbitrary shift; it requires one adapter to retain a positive expected entropy advantage. Full proofs, the entropy-boundedness argument, interpretation of the M, γ, and K dependence, and limitations of the guarantee are provided in Appendix F.

Table 1: **End-to-end CIL performance** ($\bar{A}/A_T$, %). Published rows follow the CL-LoRA protocol (He et al., 2025); reproduced CL-LoRA and CAR use identical frozen checkpoints, with CAR changing inference only.

Method	Type	CIFAR-100	ImageNet-R	ImageNet-A	VTAB
L2P (Wang et al., 2022b)	Prompt	85.50±1.23/79.51±0.67	65.82±0.71/60.62±1.12	39.81±1.36/37.62±1.89	78.96±1.62/76.41±2.26
DualPrompt (Wang et al., 2022a)	Prompt	86.96±1.98/80.44±1.38	67.41±0.30/61.73±0.93	56.43±2.33/47.45±0.96	82.51±3.49/80.94±2.87
CODA-Prompt (Smith et al., 2023)	Prompt	88.17±0.61/81.36±0.88	70.39±0.49/63.93±0.82	60.70±0.94/51.61±0.63	92.27±0.61/89.49±0.42
LAE w/ LoRA (Gao et al., 2023)	LoRA	85.17±1.53/79.67±1.06	67.55±1.22/57.04±1.13	63.25±2.21/54.28±0.94	82.24±2.45/76.00±8.21
APER (Zhou et al., 2025)	Adapter	89.09±0.56/83.26±0.52	74.05±0.30/67.13±0.63	65.53±1.16/56.60±1.81	88.27±0.16/84.99±0.06
RanPAC (McDonnell et al., 2023)	Classifier	91.63±0.28/87.62±0.16	78.53±0.73/71.06±0.71	66.14±1.54/54.85±1.36	89.61±4.21/88.85±1.36
EASE (Zhou et al., 2024)	Adapter	90.96±0.83/85.71±0.76	78.04±0.67/71.43±0.18	68.92±2.06/59.25±0.88	93.01±0.33/92.85±0.88
O-LoRA (Wang et al., 2023)	LoRA	89.63±0.61/81.26±0.68	72.52±0.29/63.19±0.26	55.02±0.74/47.53±0.84	87.22±1.21/86.98±0.89
InfLoRA (Liang & Li, 2024)	LoRA	88.84±0.90/80.97±0.74	73.22±1.12/64.51±1.25	56.91±1.27/47.04±0.90	88.83±0.94/87.16±1.17
CL-LoRA (published) (He et al., 2025)	LoRA bank	91.02±0.12/85.32±0.08	81.58±0.59/74.51±0.14	70.15±2.23/60.54±0.63	94.57±0.60/94.29±0.34
CL-LoRA (reproduced)	LoRA bank	92.52/87.94	77.83/69.48	64.97/48.78	93.80/92.77
CL-LoRA + CAR ($M = 8$)	**Context routing**	**98.74 / 98.06**	**88.97 / 82.97**	**68.81 / 52.40**	**94.95 / 93.94**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate CAR on frozen CL-LoRA checkpoints (He et al., 2025) for CIFAR-100 (Krizhevsky, 2009), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), and VTAB (Zhai et al., 2019), using a ViT-B/16 (Dosovitskiy et al., 2021) backbone pretrained on ImageNet-21K (Deng et al., 2009). The principal protocols use 10×10 classes for CIFAR-100, 40×5 for ImageNet-R, and 10×20 for ImageNet-A; VTAB follows the corresponding 10-class incremental protocol. CL-LoRA is our primary testbed because it exposes an explicit bank of task-specific LoRA modules. To test cross-method transfer, we also apply the *same frozen CAR rule* to EASE (Zhou et al., 2024) on ImageNet-R, whose task-specific adapter subspaces are learned differently. CAR never retrains either learner. Full class orders, optimization settings, seeds, checkpoint provenance, and CAR hyperparameters are in Appendix C.

For mechanism studies, same-class groups are an *analysis-only coherence control*: labels construct the group but never enter entropy computation, routing, or prediction. We then destroy this structure by shuffling the stream and ask whether useful neighborhoods can be reconstructed without labels. Appendix H distinguishes deployment-observable, analysis-only, and reconstructed contexts. We report average continual accuracy $\bar{A} = \frac{1}{T}\sum_t A_t$, final-stage accuracy A_T, and routing accuracy only as a diagnostic; task identity is never supplied to CAR.

5.2 DOES CONTEXTUAL ROUTING IMPROVE END-TO-END CIL?

Our first question is whether contextual routing improves the actual CIL objective rather than routing accuracy alone. Table 1 shows that CAR ($M = 8$) improves both $\bar{A}$ and A_T on every reproduced CL-LoRA benchmark. The largest gains are CIFAR-100 ($92.52 \rightarrow 98.74$ average; $87.94 \rightarrow 98.06$ final) and ImageNet-R ($77.83 \rightarrow 88.97$ average; $69.48 \rightarrow 82.97$ final). ImageNet-A and VTAB also improve in final accuracy ($48.78 \rightarrow 52.40$ and $92.77 \rightarrow 93.94$). Published baselines are multi-run references, whereas our reproduced checkpoint/CAR comparisons are currently single-seed; we therefore attribute gains only relative to the matched frozen checkpoint.

5.3 ANALYSIS AND DIAGNOSTICS

Does CAR generalize beyond one modular learner? We apply CAR unchanged to frozen EASE (Zhou et al., 2024) checkpoints on ImageNet-R. EASE exhibits the same monotonic context-size trend as CL-LoRA, reaching 92.38% final accuracy and 96.45% routing accuracy with large coherent contexts, compared with 71.43% native final accuracy. This supports CAR as an inference rule over specialized frozen components rather than a CL-LoRA-specific modification. The full cross-method curves are in Appendix D, Fig. 3.

How much context is needed, and why does it help? Table 2 fixes context quality with the same-class control while varying M. Moving from $M = 1$ to only two observations raises CIFAR-100 final/routing accuracy from $87.95/88.42\%$ to $94.91/95.78\%$; $M = 4$ and 8 improve further with diminishing returns. This mirrors Sec. 4.7: under a positive local entropy margin, averaging suppresses single-image routing errors.

Table 2: **Context-size progression on CIFAR-100 with CL-LoRA.** For $M > 1$, groups are formed by true class as an analysis-only coherence control; labels are not used in the CAR scoring rule.

Context M	Final acc.	Routing acc.	Δ final vs. $M = 1$
1	87.95	88.42	–
2	94.91	95.78	+6.96
4	97.53	98.52	+9.58
8	**98.06**	**99.08**	**+10.11**

A complementary stage-wise diagnostic in Appendix D, Fig. 4, shows that CAR's advantage grows as the module bank expands and sample-wise selection becomes more ambiguous.

Can CAR recover useful context when no coherent group is given? The coherent control isolates aggregation, but deployment may provide only independent or shuffled samples. With $M = 1$, minimum-entropy routing already identifies the originating task 88.42% of the time, yet final accuracy remains 87.95%, essentially the 87.94% native baseline: the adapter bank contains task information, but the single-image estimate is too noisy. We therefore test whether useful context can be reconstructed automatically. Feature kNN/DBSCAN use visual similarity, signature kNN uses adapter-response patterns, and pseudo-class grouping uses native predictions. None recovers the coherent-context gain, whereas a true-class oracle reaches 100% routing and 98.98% final accuracy. Thus, *context discovery* is the bottleneck. Full grouping and purity diagnostics are in Appendix D.

Is multiple evaluation alone sufficient? No. Eight stochastic views of one image reach only 88.67% final / 89.13% routing accuracy, versus $98.06\%/99.08\%$ for eight distinct coherent observations. The gain therefore comes from repeated evidence across distinct observations, not repeated forward passes alone; full controls are in Appendix G.

What if context is unreliable? We progressively replace coherent CIFAR-100 observations with unrelated examples. Always-on CAR remains strong through 50% corruption (97.07%) but collapses for fully random context (8.70%), while the mean routing margin falls from 0.0275 to 0.0020. With diagnostic threshold $\tau = 0.01$, gated CAR yields 87.83% under full corruption, close to the 87.94% native baseline. Thus, $\Delta(G)$ is a useful reliability diagnostic, not a formal no-harm guarantee; the full sweep is in Appendix D, Table 4.

6 Discussion and Limitations

CAR is optional: native sample-wise inference remains available when reliable context is absent. Our strongest contexts are same-class analysis controls, not evidence that arbitrary deployment streams are naturally coherent; Appendix H separates deployment-observable and analysis-only groupings, and Appendix D shows that generic reconstruction from shuffled data remains difficult. CAR also adds candidate-module inference cost, entropy is only a compatibility surrogate rather than a calibrated posterior (Guo et al., 2017; Ovadia et al., 2019), and the gating threshold is diagnostic rather than independently tuned. Future work should evaluate naturally structured streams, deployment-valid context discovery, and cheaper compatibility scoring.

7 Conclusion

We introduced Contextual Adapter Routing (CAR), a frozen context-conditioned inference strategy for modular rehearsal-free CIL. By aggregating task-local uncertainty across a small coherent unlabeled context and using it as a soft prior for all-class prediction, CAR improves multiple frozen continual learners without replay, retraining, or test-time updates. Empirical controls identify coherent multi-sample evidence as the key mechanism, and our theory shows that a positive expected entropy margin yields increasingly reliable routing as context grows. CAR therefore motivates inference context—in addition to training and memory design—as a useful axis for continual learning.

AI USE DISCLOSURE

Generative AI tools were used for research brainstorming; refining hypotheses, experimental design, and methodological framing; formulating and checking mathematical claims and proof drafts; interpreting and organizing experimental results; literature discovery; and drafting/editing manuscript text. AI tools also assisted with code- and figure-related workflows. The authors manually checked AI-assisted mathematical arguments, experimental claims, citations, code changes, and manuscript text against the underlying implementations, results, and source literature, and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

We provide the complete CAR inference rule, theoretical assumptions, and evaluation metrics in the main paper. The appendix contains full proofs of the theoretical results, expanded routing and grouping diagnostics, details of the inference-time controller study that motivated CAR, and a reproduction and provenance checklist specifying the information required to reconstruct the continual-learning checkpoints, contextual episodes, routing scores, and reported metrics. We will additionally provide anonymized implementation details and code artifacts as supplementary material where permitted by the submission system.

REFERENCES

Sk Miraj Ahmed, Fahim Faisal Niloy, Xiangyu Chang, Dripta S. Raychaudhuri, Samet Oymak, and Amit K. Roy-Chowdhury. CONTRAST: Continual multi-source adaptation to dynamic distributions. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ec4c05921d70202023898568a2a8c002-Abstract-Conference.html.

Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

Qiyuan Dai and Sibei Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online em. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9538–9548, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dai_Free_on_the_Fly_Enhancing_Flexibility_in_Test-Time_Adaptation_with_CVPR_2025_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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

Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

Takuma Fukuda, Hiroshi Kera, and Kazuhiko Kawamoto. Adapter merging with centroid prototype mapping for scalable class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4884–4893, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fukuda_Adapter_Merging_with_Centroid_Prototype_Mapping_for_Scalable_Class-Incremental_Learning_CVPR_2025_paper.html.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11483–11493, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Gao_A_Unified_Continual_Learning_Framework_with_General_Parameter-Efficient_Tuning_ICCV_2023_paper.html.

Zijian Gao, Wangwang Jia, Xingxing Zhang, Dulan Zhou, Kele Xu, Feng Dawei, Yong Dou, Xinjun Mao, and Huaimin Wang. Knowledge memorization and rumination for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20523–20533, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gao_Knowledge_Memorization_and_Rumination_for_Pre-trained_Model-based_Class-Incremental_Learning_CVPR_2025_paper.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. CL-LoRA: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30534–30544, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/He_CL-LoRA_Continual_Low-Rank_Adaptation_for_Rehearsal-Free_Class-Incremental_Learning_CVPR_2025_paper.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Jiashuo Li, Shaokun Wang, Bo Qian, Yuhang He, Xing Wei, Qiang Wang, and Yihong Gong. Dynamic integration of task-specific adapters for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30545–30555, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Dynamic_Integration_of_Task-Specific_Adapters_for_Class_Incremental_Learning_CVPR_2025_paper.html.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liang_InfLoRA_Interference-Free_Low-Rank_Adaptation_for_Continual_Learning_CVPR_2024_paper.html.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2793dc35e14003dd367684d93d236847-Abstract-Conference.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Mincheol Park, Heeji Won, Won Woo Ro, and Suhyun Kim. Rethinking entropy in test-time adaptation: The missing piece from energy duality. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/435422305988b73c6cc00bcb29ba2531-Abstract-Conference.html.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Rebuffi_iCaRL_Incremental_Classifier_CVPR_2017_paper.html.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. CODA-Prompt: COntinual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Smith_CODA-Prompt_COntinual_Decomposed_Attention-Based_Prompting_for_Rehearsal-Free_Continual_Learning_CVPR_2023_paper.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Huiyi Wang, Haodong Lu, Lina Yao, and Dong Gong. Self-expansion of pre-trained models with mixture of adapters for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10087–10098, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Self-Expansion_of_Pre-trained_Models_with_Mixture_of_Adapters_for_Continual_CVPR_2025_paper.html.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023. URL https://aclanthology.org/2023.findings-emnlp.715/.

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*, pp. 806–816, 2025b. doi: 10.1109/ICCV51701.2025.00083. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Integrating_Task-Specific_and_Universal_Adapters_for_Pre-Trained_Model-based_Class-Incremental_Learning_ICCV_2025_paper.html.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2022a. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/5428_ECCV_2022_paper.php.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Learning_To_Prompt_for_Continual_Learning_CVPR_2022_paper.html.

Zixin Wang, Dong Gong, Sen Wang, Zi Huang, and Yadan Luo. Is less more? exploring token condensation as training-free test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 144–154, 2025c. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Is_Less_More_Exploring_Token_Condensation_as_Training-free_Test-time_Adaptation_ICCV_2025_paper.html.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. SD-LoRA: Scalable decoupled low-rank adaptation for class incremental learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5U1rlpX68A.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. URL https://arxiv.org/abs/1910.04867.

Binyu Zhao, Wei Zhang, Xingrui Yu, Zhaonian Zou, and Ivor Tsang. Advancing analytic class-incremental learning through vision-language calibration. *arXiv preprint arXiv:2602.13670*, 2026. URL https://arxiv.org/abs/2602.13670.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido M. van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/47197750b407ef48daa1438e0c0af242-Abstract-Conference.html.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23554–23564, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_Expandable_Subspace_Ensemble_for_Pre-Trained_Model-Based_Class-Incremental_Learning_CVPR_2024_paper.html.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025. doi: 10.1007/s11263-024-02218-0. URL https://doi.org/10.1007/s11263-024-02218-0.

Algorithm 1 Contextual Adapter Routing (CAR)

Require: Context $G = \{x_i\}_{i=1}^{M}$, adapter bank $\{A_k\}_{k=1}^{K}$, task class sets $\{\mathcal{C}_k\}_{k=1}^{K}$

for $k = 1, \ldots, K$ **do**
  for $i = 1, \ldots, M$ **do**
    Compute $p_{k,c}(x_i)$ using Eq. (3)
    Compute $h_k(x_i)$ using Eq. (4)
  end for
  $H_k(G) \leftarrow \frac{1}{M} \sum_{i=1}^{M} h_k(x_i)$
end for
$\hat{k}(G) \leftarrow \arg\min_k H_k(G)$
Construct $\{\pi_k(G)\}_{k=1}^{K}$ using Eq. (8)
for $i = 1, \ldots, M$ **do**
  Compute $s_c(x_i \mid G)$ for all $c \in \mathcal{C}_{\leq t}$
  $\hat{y}_i \leftarrow \arg\max_{c \in \mathcal{C}_{\leq t}} s_c(x_i \mid G)$
end for
return $\{\hat{y}_i\}_{i=1}^{M}$

Table 3: **Inference capabilities of representative rehearsal-free PTM-based CIL methods.** A checkmark denotes that the method satisfies the corresponding property. The table summarizes representative inference properties; CAR adds cross-sample contextual inference to a frozen modular learner.

Method	Rehearsal-free	PTM-based	Task-specialized components	Task-ID-free inference	Cross-sample contextual inference
Shared / non-expanding continual adaptation					
EWC-LoRA (Zheng et al., 2026)	✓	✓	✗	✓	✗
RanPAC (McDonnell et al., 2023)	✓	✓	✗	✓	✗
SimpleCIL (Zhou et al., 2025)	✓	✓	✗	✓	✗
VILA (Zhao et al., 2026)	✓	✓	✗	✓	✗
Prompt / expandable task-conditioned continual adaptation					
L2P (Wang et al., 2022b)	✓	✓	✓	✓	✗
DualPrompt (Wang et al., 2022a)	✓	✓	✓	✓	✗
CODA-Prompt (Smith et al., 2023)	✓	✓	✓	✓	✗
EASE (Zhou et al., 2024)	✓	✓	✓	✓	✗
CL-LoRA (He et al., 2025)	✓	✓	✓	✓	✗
CAR (ours)	✓	✓	✓	✓	✓

A Complete CAR Algorithm

Inference cost. For a context of size M and K candidate adapters, contextual scoring requires task-local responses for all KM adapter–observation pairs. In our CL-LoRA implementation, a single model call per image batch evaluates the shared backbone and all currently seen adapters, after which the task-local entropies and native all-class logits are cached. CAR therefore does not require a second selected-adapter forward pass. It performs no backward pass, optimization step, parameter update, or replay operation.

B Inference-Setting Comparison

C Reproduction and Provenance

This appendix records the information required to reproduce both the underlying continual-learning checkpoints and the proposed inference-time analyses. A central distinction throughout the paper is whether an experimental grouping rule uses only information that would be available at deployment or whether it is an analysis-only coherence control. We therefore recommend archiving not only model and optimization state, but also the exact construction of every test context.

C.1 Continual-learning configuration

For each benchmark and continual learner, the following information should be archived:

- dataset version and preprocessing pipeline;

- train, validation, and test split identifiers;
- class order and task partition;
- number of continual stages T;
- classes assigned to each $\mathcal{C}_t$;
- backbone architecture and pretrained checkpoint;
- continual-learning method and all method-specific hyperparameters;
- optimizer, learning-rate schedule, number of epochs, and batch size;
- random seed;
- checkpoints saved after every continual stage;
- code version or commit hash used to train each checkpoint.

Because CAR does not alter continual training, the most important provenance requirement is that the native baseline and CAR are evaluated using *exactly the same frozen checkpoint*.

CIFAR-100 diagnostic configuration. The principal ten-task diagnostic uses seed 1993, CIFAR-100 split into 10 tasks of 10 classes, and a ViT-B/16 (Dosovitskiy et al., 2021) ImageNet-21K-pretrained (Deng et al., 2009) CL-LoRA backbone. The LoRA bottleneck dimension is 8 and LoRA is inserted into the query and value projections. Blocks 0–5 are shared and blocks 6–11 are task-specific. Training uses SGD with learning rate 0.03, momentum 0.9, weight decay 10^{-4}, cosine decay to zero, 30 epochs per task, and batch size 64. The paired routing diagnostic is reproduced by

```
python tools/run_routing_control.py exps/cifar_t10.json
-device 0 -seed 1993 -eval-batch-size 64 -shuffle-seed 2026
```

The stream/view seed 2026 controls evaluation ordering or augmentations and is not an additional training seed.

Other benchmark configurations. For CL-LoRA, ImageNet-R uses 40 tasks of 5 classes, bottleneck dimension 10, SGD learning rate 0.05, 20 epochs per task, batch size 32, and weight decay 5×10^{-4}. ImageNet-A uses 10 tasks of 20 classes, learning rate 0.05, 25 epochs per task, batch size 32, and weight decay 5×10^{-3}. VTAB uses 10-class increments, learning rate 0.03, 45 epochs per task, batch size 32, and weight decay 10^{-4}. The EASE ImageNet-R generalization experiment uses seed 1993, 40 tasks of 5 classes, batch size 16, learning rate 0.05, 20 epochs per task, and weight decay 5×10^{-3}.

C.2 CAR INFERENCE CONFIGURATION

For every CAR evaluation, the following quantities should be stored:

- context size M;
- candidate adapter set $\mathcal{A}_{\leq t}$;
- class block $\mathcal{C}_k$ associated with each adapter;
- exact logits used for task-local entropy computation;
- raw task-local entropy is used (no normalization by $\log |\mathcal{C}_k|$);
- no additional routing temperature is applied ($T = 1$);
- smoothing parameter $\epsilon = 10^{-3}$;
- contextual-prior strength $\lambda = 1$;
- the exact rule used to construct each context G;
- image identifiers and ordering within every context;
- per-image adapter entropies $\{h_k(x_i)\}$;
- contextual scores $\{H_k(G)\}$;
- selected adapter $\hat{k}(G)$;
- final class prediction for each observation;
- routing correctness, when ground-truth task identity is available for evaluation;
- number of candidate-module evaluations;
- wall-clock inference latency and hardware configuration.

For each context-construction procedure, we additionally record whether the group can be formed using information observable at deployment. In particular, same-class and true-task grouping are

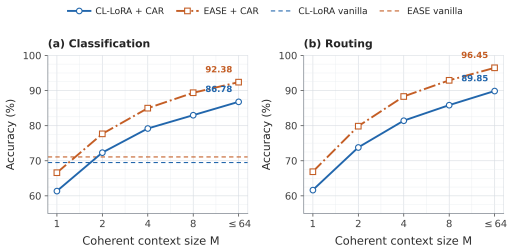


Figure 3: **Cross-method generalization on ImageNet-R.** Coherent context improves final classification and routing for both CL-LoRA and EASE; dashed lines show native final accuracy.

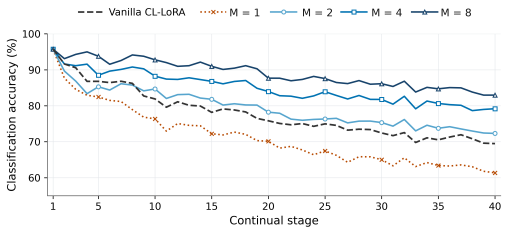


Figure 4: **Stage-wise ImageNet-R diagnostic.** As task-specialized modules accumulate, contextual routing increasingly outperforms sample-wise entropy routing.

used only as controlled analyses of context coherence and are never presented as deployment-valid grouping mechanisms.

C.3 Evaluation quantities

Let A_t denote all-class classification accuracy after continual stage t. The two primary CIL metrics are

$$\bar{A} = \frac{1}{T}\sum_{t=1}^{T} A_t, \qquad A_T, \tag{18}$$

where $\bar{A}$ is average continual accuracy and A_T is final-stage accuracy.

For routing diagnostics, let $k_i^\star$ denote the ground-truth originating task of test sample x_i, used *only for evaluation*. Routing accuracy is

$$A_{\text{route}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{k}(G_i) = k_i^\star\right]. \tag{19}$$

Task identities in Eq. (19) are never supplied to the routing algorithm.

D Expanded Routing and Grouping Diagnostics

The main paper reports only a compact subset of the routing diagnostics. Here we retain the broader sweep, including unsuccessful grouping strategies, because these results help identify what CAR does and does not exploit.

The central question is whether the strong performance of CAR can be recovered when the test stream is globally shuffled and context coherence must itself be estimated. The answer in the current experiments is largely negative: several feature-space, adapter-response, and pseudo-label grouping approaches fail to recover the gains obtained from explicitly coherent contexts.

D.1 Context-corruption robustness

The full CIFAR-100 corruption sweep used in Sec. 5.3 is reported below. Corruption labels are used only to construct the diagnostic contexts.

D.2 Expanded diagnostic results

D.3 Single-image minimum-entropy routing

The simplest diagnostic selects an adapter independently for each image:

Table 4: **Robustness to context corruption on CIFAR-100.** Final-stage accuracy for $M = 64$. The confidence gate uses diagnostic threshold $\tau = 0.01$; no separate validation split was used to select it.

Corruption	Native	CAR	Gated CAR	Mean $\Delta(G)$
0%	87.94	98.62	97.18	0.0275
12.5%	87.94	98.86	96.53	0.0241
25%	87.94	98.31	95.69	0.0203
50%	87.94	97.07	92.65	0.0128
75%	87.94	86.87	88.42	0.0053
100%	87.94	8.70	87.83	0.0020

Table 5: **Expanded ten-task CIFAR-100 routing diagnostics.** Final and routing accuracies are percentages. Rows based on true-class grouping are label-informed analysis controls; shuffled grouping methods do not use labels to construct neighborhoods. Different rows use different neighborhood structures, so this is not a controlled context-size ablation.

Family	Method / grouping rule	Context construction	Size / key setting	Final acc.	Route acc.
No routing	Vanilla CL-LoRA	Native	1	87.94	–
Single image	Minimum-entropy adapter	Native	1	87.95	88.42
Ordered diagnostic	Consecutive cached samples	Ordered stream	$M = 64$	97.32	98.32
Oracle	True-class grouping	Label-informed	$M = 100$	98.98	100.00
Unsupervised grouping	DBSCAN	Shuffled features	$\varepsilon = 0.3$	74.26	74.66
Unsupervised grouping	Feature kNN	Shuffled features	$k = 5$	78.85	79.26
Adapter-response grouping	Signature kNN	Shuffled signatures	$k = 5$	86.60	86.97
Adapter-response grouping	Signature kNN	Shuffled signatures	$k = 10$	86.76	87.09
Adapter-response grouping	Signature kNN	Shuffled signatures	$k = 20$	86.54	86.86
Pseudo-class grouping	Hard pseudo-class	Shuffled	variable	87.94	88.19
Pseudo-class grouping	High-confidence (≥ 0.8)	Shuffled	variable	87.94	–[†]
Pseudo-class grouping	Iterative, budgets 1–10	Shuffled	variable	87.94	88.19

[†] At threshold 0.8, no sample qualified as a confidence estimator in the final-stage run (coverage 0%); predictions therefore fall back to the native classifier and routing accuracy is undefined.

$$\hat{k}(x) = \arg\min_k h_k(x). \tag{20}$$

This already yields 88.42% routing accuracy in the ten-task pilot, showing that task-specialized modules contain substantial task-discriminative information even though they were not explicitly trained as task classifiers. However, final all-class accuracy changes only from 87.94% to 87.95%. Thus, the per-image signal is informative but insufficiently reliable to produce a substantial end-to-end benefit.

D.4 Ordered coherent batches

The ordered-batch experiment uses the cached cumulative test order with `shuffle=False` and splits consecutive samples into groups of size 64 (retaining the final partial group). The cumulative test set is assembled from class blocks, so this ordering contains strong local structure, although a 64-sample group need not be perfectly class-homogeneous. This analysis yields 98.32% routing accuracy and 97.32% final accuracy.

Because the stored ordering itself provides substantial structure, this experiment is treated as a diagnostic rather than a general deployment grouping rule. Its purpose is to show that repeated correlated task evidence can make entropy aggregation much more reliable than single-image routing.

D.5 True-class oracle grouping

The strongest diagnostic constructs groups using ground-truth class identity. For each class c, observations satisfying $y_i = c$ are grouped before routing. The group labels are not used in the entropy score or the final class prediction.

This analysis reaches 100% routing accuracy and 98.98% final accuracy. It therefore provides an approximate upper-bound diagnostic on what can happen when the context is maximally semantically coherent.

The oracle is not a deployment algorithm. Instead, it demonstrates that the remaining difficulty lies largely in obtaining a sufficiently coherent unlabeled context rather than in designing a substantially more complicated adapter-selection rule.

D.6 FEATURE-SPACE GROUPING

We also test whether coherence can be reconstructed directly from frozen image features using nearest-neighbor neighborhoods (Cover & Hart, 1967).

Let $\phi(x) \in \mathbb{R}^d$ denote the frozen adapter-free pretrained representation used for grouping. We ℓ_2-normalize these features and construct exact cosine-similarity neighborhoods within the current cumulative test pool. The query itself is excluded, and deterministic ties are resolved by stable sample identifier. CAR then averages the *raw adapter entropies* of the k neighboring samples and routes the query from that neighborhood evidence.

In the shuffled ten-task pilot, feature kNN with $k = 5$ obtains 79.26% routing accuracy and 78.85% final accuracy, substantially below both native inference and explicitly coherent contexts. Sweeps with $k = 10, 20, 50$ similarly fail to recover the coherent-context gain.

This suggests that generic feature proximity is not sufficient to reconstruct the form of semantic coherence required by CAR.

D.7 DBSCAN GROUPING

We additionally apply DBSCAN (Ester et al., 1996), a density-based clustering algorithm, to the shuffled feature stream. Features are frozen adapter-free representations, normalized before clustering. DBSCAN uses cosine distance, `min_samples`=5, and is refit on the current cumulative test pool at each stage. The completed cached $\varepsilon = 0.3$ comparison obtains 74.66% routing accuracy and 74.26% final classification accuracy. Cluster members are processed in shuffled arrival order in chunks of at most 64; noise points are routed as singletons.

This diagnostic does not maintain persistent online cluster centers across continual stages. It should therefore be interpreted as an offline/transductive clustering baseline rather than an online context-discovery mechanism.

D.8 ADAPTER-RESPONSE SIGNATURES

Because the routing signal originates from the adapter bank itself, we also construct neighborhoods directly in adapter-response space. For each seen adapter k, we extract three task-local statistics: predictive entropy $h_k(x)$, maximum task-local probability, and the margin between the two largest task-local probabilities. Concatenating these quantities over K adapters gives

$$r(x) = [h_k(x),\ p_k^{\max}(x),\ m_k(x)]_{k=1}^{K} \in \mathbb{R}^{3K}. \tag{21}$$

Each signature coordinate is standardized over the current test pool using its population mean and standard deviation; numerically constant coordinates are set to zero. We then use exact cosine kNN in this standardized signature space, exclude the query itself, and average the *original unstandardized adapter entropies* of the selected neighbors to obtain the query's CAR route.

The resulting final/routing accuracies are

$$k = 5: \quad 86.60/86.97, \tag{22}$$
$$k = 10: \quad 86.76/87.09, \tag{23}$$
$$k = 20: \quad 86.54/86.86, \tag{24}$$
$$k = 50: \quad 86.09/86.39. \tag{25}$$

These results are stronger than generic feature-space kNN but still fail to recover the large gains observed when genuine coherence is available. The adapter bank therefore defines a task-sensitive response representation, yet local nearest-neighbor grouping in that representation does not by itself solve context discovery.

D.9 PSEUDO-CLASS GROUPING

A different strategy is to group observations according to their current predicted class.

Let

$$\tilde{y}(x) = \arg\max_{c \in \mathcal{C}_{\leq t}} z_c^{\text{all}}(x) \tag{26}$$

denote the native pseudo-label. Hard pseudo-class grouping forms contexts from observations with the same value of $\tilde{y}$.

This procedure achieves 88.19% routing accuracy and 87.94% final accuracy, effectively recovering the native baseline but not the contextual gain.

We further evaluate a high-confidence variant that retains only predictions whose maximum class probability satisfies

$$\max_c p_c^{\text{all}}(x) \geq 0.8. \tag{27}$$

At the chosen threshold, the final-stage run has 0% estimator coverage (the largest native max-softmax confidence is only about 0.0249), so all samples fall back to the native prediction. The resulting final accuracy is therefore 87.94%, and routing accuracy is undefined for this condition.

Finally, we evaluate iterative pseudo-class grouping with update budgets from 1 to 10. In the observed runs, the procedure converges after a single effective pass and again yields 87.94% final accuracy and 88.19% routing accuracy.

These experiments indicate an important circularity: if the native classifier is already uncertain or biased toward an incorrect class, grouping by its own pseudo-label does not create sufficiently cleaner evidence to correct the routing decision.

D.10 INTERPRETATION

Taken together, the expanded diagnostics reveal a clear hierarchy.

Single-image entropy demonstrates that the adapter bank contains a useful latent routing signal. Generic feature clustering performs poorly. Grouping in adapter-response space preserves more task information but still remains insufficient. Pseudo-label grouping similarly fails to bootstrap a stronger context from globally shuffled observations. In contrast, ordered, same-class, and oracle contexts produce very high routing reliability.

The primary bottleneck is therefore not the absence of a sophisticated adapter-scoring rule. Rather, it is whether the available observations provide multiple approximately independent measurements of the same underlying adapter preference.

E DISCOVERY THROUGH INFERENCE-TIME POLICY LEARNING

CAR emerged from a broader inference-controller study. We include this experiment only as a discovery diagnostic: the learned policy strongly concentrated on minimum-entropy adapter selection and provided no material advantage over the corresponding deterministic heuristic.

E.1 CONTROLLER STATE AND ACTION SPACE

The historical controller operates once per test batch ($B = 64$) and receives an eight-dimensional state

$$s = \left[\bar{H}_{\text{all}}, \bar{p}_{\max}, \text{Std}(p_{\max}), c_{\max}, d_{\text{adapt}}, \overline{\|f(x)\|_2}, \rho_{\text{proto}}, \mathbf{1}[\bar{p}_{\max} \geq 0.6]\right]^\top, \tag{28}$$

where $\bar{H}_{\text{all}}$ is mean entropy of the full-class prediction, $\bar{p}_{\max}$ is mean maximum softmax confidence, $c_{\max}$ is the largest candidate consistency score, d_{adapt} measures disagreement between adapter-

conditioned full-class probabilities, and ρ_{proto} is the mean maximum cosine similarity to available class prototypes.

The seven-action vocabulary is

$$\mathcal{U} = \{\texttt{default}, \texttt{lowest_entropy_adapter}, \texttt{ensemble_adapters}, \texttt{entropy_update}, \texttt{consistency_update}, \texttt{skip_or_reset}, \texttt{entropy_weighted_average}\}. \tag{29}$$

The two update actions can perform one LoRA-only test-time update with learning rate 10^{-4}; the experiment resets these updates after each batch, and the logged base-weight update count is zero. Thus this historical controller is strictly broader than the final CAR method, which performs no test-time update.

E.2 Policy architecture and online learning rule

The controller is a two-layer MLP,

$$8 \to 32 \to 32,$$

with ReLU nonlinearities, a seven-way policy head, and a scalar value head. Actions are sampled from the categorical policy during the online experiment. The controller uses Adam with learning rate 10^{-3} and gradient-norm clipping at 5.

This experiment is best described as an online contextual-bandit/actor–critic update rather than PPO. Let r denote the immediate shaped reward and b the running reward baseline. The implementation updates

$$b \leftarrow 0.9b + 0.1r, \qquad A = r - b, \tag{30}$$

and minimizes

$$\mathcal{L} = -\log \pi_\psi(a \mid s)\ \text{stopgrad}(A) + \big(V_\psi(s) - r\big)^2 - 0.01\,\mathcal{H}(\pi_\psi(\cdot \mid s)). \tag{31}$$

There is no discounted rollout or replay-buffer return in this update.

E.3 Reward

The reward is label-free:

$$r = -H_{\text{post}} + C_{\text{cons}} - 0.1\|\Delta\theta\|_2 - c(a), \tag{32}$$

where H_{post} is post-action predictive entropy, C_{cons} is an augmentation-consistency score based on the symmetric KL divergence between the original and horizontally flipped predictions, $\|\Delta\theta\|_2$ is the test-time update norm, and

$$c(a) \in \{0, 0.02, 0.04, 0.06, 0.08, 0.01, 0.03\}$$

for the seven actions in the order listed above. Labels are not used in the reward.

E.4 Observed policy concentration

In the completed seven-action CIFAR-100 run, the controller chose `lowest_entropy_adapter` on 835 of 864 decisions (96.64%). The remaining decisions were: ensemble 12, consistency update 8, skip/reset 3, entropy-weighted average 3, default 2, and entropy update 1. The final 100 decisions all selected the minimum-entropy action; the saved final policy probability for this action is approximately 0.965.

The full learned controller reaches 97.490% average continual accuracy and 96.62% final accuracy in this historical run. A separately executed deterministic run restricted to the lowest-entropy action reaches 97.537% average accuracy and the same 96.62% final accuracy. Because these are historical separately executed runs rather than a paired shared-checkpoint comparison, we use them only to document the discovery process.

A later paired per-sample cached experiment gives the same qualitative result: the learned four-action controller obtains 92.498% average and 87.95% final accuracy, identical in final accuracy to deterministic hard entropy routing; on the final stage it chooses the hard entropy action for 9993 of 10000 samples.

Table 6: **Action counts in the completed seven-action controller run.**

Action	Count
Lowest-entropy adapter	835
Ensemble adapters	12
Consistency update	8
Skip/reset	3
Entropy-weighted average	3
Default	2
Entropy update	1

E.5 From controller behavior to CAR

The controller study suggested that useful task information was already present in task-local entropy, but that single-sample selection remained noisy. CAR therefore removes the learned policy entirely and changes the estimator rather than increasing policy complexity:

$$H_k(G) = \frac{1}{M}\sum_{i=1}^{M} h_k(x_i), \qquad \hat{k}(G) = \arg\min_k H_k(G). \tag{33}$$

The final method has no policy network, no controller objective, and no test-time parameter update. The controller is retained here only to document how the minimum-entropy signal was discovered.

F Proofs for Contextual Routing

We now provide the full proof of the theoretical statements in the main paper.

F.1 Setup

Let Q denote the local test distribution generating the contextual episode

$$G = \{x_1, \dots, x_M\}, \qquad x_i \overset{\text{i.i.d.}}{\sim} Q. \tag{34}$$

Importantly, Q need not equal any distribution encountered during continual training. It may represent a shifted local test distribution.

For candidate adapter k, define the expected task-local entropy under Q as

$$\mu_k(Q) = \mathbb{E}_{X\sim Q}[h_k(X)]. \tag{35}$$

We define the population-optimal adapter for Q as

$$k^\star = \arg\min_{k\in\{1,\dots,K\}} \mu_k(Q). \tag{36}$$

For each competing adapter $j \neq k^\star$, define

$$D_j(X) = h_j(X) - h_{k^\star}(X). \tag{37}$$

The expected entropy margin is

$$\mu_j := \mathbb{E}_{X\sim Q}[D_j(X)] = \mu_j(Q) - \mu_{k^\star}(Q). \tag{38}$$

The theorem assumes that this quantity is bounded away from zero:

$$\mu_j \geq \gamma_j > 0. \tag{39}$$

Thus, the correct adapter need not have lower entropy on every individual sample. It only needs to have lower entropy *in expectation* under the current local distribution.

F.2 Main theorem

Theorem 2 (Reliable contextual routing under a local entropy margin)**.** *Let* $G = \{x_i\}_{i=1}^{M}$ *with* $x_i \overset{\text{i.i.d.}}{\sim} Q$*. Let* $k^\star$ *be defined by Eq.* (36)*.*

For every competitor $j \neq k^\star$*, assume that:*

1. $D_j(X)$ *is almost surely contained in an interval* $[a_j, b_j]$ *of width*
$$R_j = b_j - a_j;$$
2. *there exists* $\gamma_j > 0$ *such that*
$$\mathbb{E}_{X\sim Q}[D_j(X)] \geq \gamma_j.$$

Then the CAR routing rule

$$\hat{k}(G) = \arg\min_k \frac{1}{M}\sum_{i=1}^{M} h_k(x_i) \tag{40}$$

satisfies

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - \sum_{j\neq k^\star} \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right). \tag{41}$$

F.3 Proof of Theorem 2

Proof. For any candidate adapter k, CAR computes

$$H_k(G) = \frac{1}{M}\sum_{i=1}^{M} h_k(x_i). \tag{42}$$

Fix one competitor $j \neq k^\star$.

CAR prefers j to $k^\star$ whenever

$$H_j(G) \leq H_{k^\star}(G). \tag{43}$$

Substituting Eq. (42),

$$\frac{1}{M}\sum_{i=1}^{M} h_j(x_i) \leq \frac{1}{M}\sum_{i=1}^{M} h_{k^\star}(x_i). \tag{44}$$

Rearranging,

$$\frac{1}{M}\sum_{i=1}^{M} (h_j(x_i) - h_{k^\star}(x_i)) \leq 0. \tag{45}$$

Using the definition $D_j(x) = h_j(x) - h_{k^\star}(x)$,

$$H_j(G) \le H_{k^\star}(G) \quad \Longleftrightarrow \quad \frac{1}{M}\sum_{i=1}^{M} D_j(x_i) \le 0. \tag{46}$$

Define the empirical gap

$$\widehat{\mu}_j = \frac{1}{M}\sum_{i=1}^{M} D_j(x_i). \tag{47}$$

Because the samples are i.i.d.,

$$\mathbb{E}[\widehat{\mu}_j] = \mathbb{E}_{X\sim Q}[D_j(X)] = \mu_j \ge \gamma_j. \tag{48}$$

The pairwise routing error requires $\widehat{\mu}_j \le 0$. Therefore,

$$\Pr[H_j(G) \le H_{k^\star}(G)] = \Pr[\widehat{\mu}_j \le 0] \tag{49}$$

$$= \Pr[\widehat{\mu}_j - \mu_j \le -\mu_j]. \tag{50}$$

Since $\mu_j \ge \gamma_j$,

$$\{\widehat{\mu}_j - \mu_j \le -\mu_j\} \subseteq \{\widehat{\mu}_j - \mu_j \le -\gamma_j\}. \tag{51}$$

Hence,

$$\Pr[H_j(G) \le H_{k^\star}(G)] \le \Pr[\widehat{\mu}_j - \mu_j \le -\gamma_j]. \tag{52}$$

By assumption, $D_j(X) \in [a_j, b_j]$ almost surely and $R_j = b_j - a_j$. Hoeffding's inequality (Hoeffding, 1963) therefore implies

$$\Pr[\widehat{\mu}_j - \mu_j \le -\gamma_j] \le \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right). \tag{53}$$

Thus the probability that any particular competitor j defeats $k^\star$ is bounded by

$$\Pr[H_j(G) \le H_{k^\star}(G)] \le \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right). \tag{54}$$

CAR makes an incorrect routing decision only if at least one competitor achieves a contextual score no larger than that of $k^\star$:

$$\{\hat{k}(G) \ne k^\star\} \subseteq \bigcup_{j\ne k^\star} \{H_j(G) \le H_{k^\star}(G)\}. \tag{55}$$

Applying the union bound,

$$\Pr[\hat{k}(G) \ne k^\star] \le \sum_{j\ne k^\star} \Pr[H_j(G) \le H_{k^\star}(G)] \tag{56}$$

$$\le \sum_{j\ne k^\star} \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right). \tag{57}$$

Taking complements gives

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - \sum_{j \neq k^\star} \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right), \tag{58}$$

which proves the theorem.

□

F.4 Sample-complexity corollary

Corollary 2 (Context size sufficient for reliable routing)**.** *Under the assumptions of Theorem 2, define*

$$\gamma = \min_{j \neq k^\star} \gamma_j, \qquad R = \max_{j \neq k^\star} R_j. \tag{59}$$

Then

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - (K-1)\exp\left(-\frac{2M\gamma^2}{R^2}\right). \tag{60}$$

Consequently, for any target failure probability $\delta \in (0,1)$, it is sufficient that

$$M \geq \frac{R^2}{2\gamma^2}\log\left(\frac{K-1}{\delta}\right) \tag{61}$$

to guarantee

$$\Pr[\hat{k}(G) = k^\star] \geq 1 - \delta. \tag{62}$$

F.5 Proof of Corollary 2

Proof. From the definitions of γ and R,

$$\gamma_j \geq \gamma \qquad \text{and} \qquad R_j \leq R \tag{63}$$

for all $j \neq k^\star$. Therefore,

$$\frac{\gamma_j^2}{R_j^2} \geq \frac{\gamma^2}{R^2}. \tag{64}$$

Applying this inequality to each term of Eq. (41),

$$\Pr[\hat{k}(G) \neq k^\star] \leq \sum_{j \neq k^\star} \exp\left(-\frac{2M\gamma_j^2}{R_j^2}\right) \tag{65}$$

$$\leq (K-1)\exp\left(-\frac{2M\gamma^2}{R^2}\right). \tag{66}$$

To make the failure probability at most δ, it is sufficient to require

$$(K-1)\exp\left(-\frac{2M\gamma^2}{R^2}\right) \leq \delta. \tag{67}$$

Dividing by $K-1$,

$$\exp\left(-\frac{2M\gamma^2}{R^2}\right) \leq \frac{\delta}{K-1}. \tag{68}$$

Taking logarithms,

$$-\frac{2M\gamma^2}{R^2} \leq \log\left(\frac{\delta}{K-1}\right). \tag{69}$$

Multiplying by -1 reverses the inequality:

$$\frac{2M\gamma^2}{R^2} \geq \log\left(\frac{K-1}{\delta}\right). \tag{70}$$

Solving for M yields

$$M \geq \frac{R^2}{2\gamma^2}\log\left(\frac{K-1}{\delta}\right). \tag{71}$$

Substitution into the preceding failure bound gives

$$\Pr[\hat{k}(G) \neq k^\star] \leq \delta, \tag{72}$$

or equivalently

$$\Pr[\hat{k}(G) = k^\star] \geq 1-\delta. \tag{73}$$

□

F.6 Interpretation of the bound

The bound isolates three quantities controlling contextual routing.

Context size. The failure probability decreases exponentially in M:

$$P_{\text{fail}} = O\left(e^{-M\gamma^2/R^2}\right). \tag{74}$$

Thus, repeated observations from the same local distribution can convert a weak single-image signal into a substantially more reliable estimate.

Adapter separation. The required context size scales as

$$M = O\left(\frac{1}{\gamma^2}\right). \tag{75}$$

When the correct adapter has a large expected entropy advantage, very few observations are required. When γ is close to zero, substantially more context is needed.

Number of candidate adapters. The dependence on the adapter-bank size is only logarithmic:

$$M = O(\log K). \tag{76}$$

Thus, under the stated assumptions, maintaining more candidate modules does not require context size to grow linearly with the number of modules.

Distribution shift. The theorem does not assume that Q equals the training distribution. The entire argument is stated under the current test distribution Q. Therefore, distribution shift is allowed provided that the appropriate adapter preserves a positive expected entropy advantage:

$$\mathbb{E}_{X\sim Q}\left[h_j(X) - h_{k^\star}(X)\right] > 0. \tag{77}$$

The theorem therefore concerns *local ranking stability under shift*, not absence of distribution shift.

F.7 Why the boundedness assumption is mild for entropy

Predictive entropy for adapter k satisfies

$$0 \le h_k(X) \le \log|\mathcal{C}_k|. \tag{78}$$

Hence,

$$D_j(X) = h_j(X) - h_{k^\star}(X) \tag{79}$$

lies within

$$-\log|\mathcal{C}_{k^\star}| \le D_j(X) \le \log|\mathcal{C}_j|. \tag{80}$$

A valid interval width is therefore

$$R_j \le \log|\mathcal{C}_j| + \log|\mathcal{C}_{k^\star}|. \tag{81}$$

If all task blocks contain C classes, this simplifies to

$$R_j \le 2\log C. \tag{82}$$

If normalized entropy is used,

$$\widetilde{h}_k(X) = \frac{h_k(X)}{\log|\mathcal{C}_k|}, \tag{83}$$

then

$$0 \le \widetilde{h}_k(X) \le 1, \tag{84}$$

and consequently the normalized gap

$$\widetilde{D}_j(X) = \widetilde{h}_j(X) - \widetilde{h}_{k^\star}(X) \tag{85}$$

lies in $[-1, 1]$, so one may take

$$R_j \le 2. \tag{86}$$

This observation makes the concentration assumption particularly natural for entropy-based routing.

Table 7: **Same-image augmentation versus distinct coherent context.** Repeated stochastic views of one image provide only modest gains, whereas distinct coherent observations strongly improve both final CIL accuracy and routing accuracy.

Context source	M	Final acc.	Route acc.
Single image	1	87.95	88.42
Same-image augmentations	2	88.24	88.74
Same-image augmentations	4	88.48	88.94
Same-image augmentations	8	88.67	89.13
Coherent images[†]	2	94.91	95.78
Coherent images[†]	4	97.53	98.52
Coherent images[†]	8	98.06	99.08

[†] Same-class grouping is used only as an analysis-time coherence control; labels are not used by CAR for scoring or prediction.

F.8 What the theorem does not claim

The result is intentionally conditional. It does not claim that the correct adapter must always have minimum entropy, nor that arbitrary groups of test samples are useful.

If for some competitor j,

$$\mathbb{E}_{X\sim Q}[D_j(X)] \le 0, \tag{87}$$

then the expected entropy ordering is absent or reversed. In that regime, increasing M cannot be guaranteed to recover the desired adapter; indeed, aggregation may make the incorrect preference more stable.

Similarly, the i.i.d. assumption models a coherent local episode. If a constructed group instead mixes observations from substantially different distributions,

$$x_i \sim Q_i, \qquad Q_i \neq Q_j, \tag{88}$$

then the simple concentration statement above no longer applies directly.

These failure modes correspond closely to the empirical diagnostics: coherent groups yield strong routing gains, whereas arbitrary shuffled groups do not.

G Same-Image Augmentation Control

The main paper compares repeated stochastic views of the same image against multiple distinct coherent observations. We restate the experiment here because it separates contextual evidence from generic test-time ensembling.

The completed control uses the original view plus weak stochastic views generated with random-resized crop (scale $[0.9, 1.0]$, aspect ratio $[0.95, 1.05]$), horizontal flip when enabled by the training transform, and mild color jitter (brightness/contrast/saturation 0.1, hue 0.02), followed by the dataset normalization. View generation is deterministic under seed $2026 + 100000\,t + \mathrm{sampleID}$. CAR biases the *original-view* all-class logits after aggregating routing evidence; it does not average final class predictions across views. We report mean per-view entropy below. An alternative entropy-of-mean-probability aggregation gives nearly identical $M = 8$ performance (88.69% final, 89.15% routing).

If repeated forward passes alone explained CAR, the augmentation and multi-image curves should improve comparably with M. Instead, eight same-image views improve final accuracy by less than one percentage point over single-image entropy routing, whereas eight distinct coherent observations produce a gain of more than ten percentage points in the same diagnostic setting. This strongly suggests that the relevant quantity is repeated evidence across distinct observations rather than simple variance reduction from repeatedly processing one input.

H Additional Notes on Context Construction

The current experiments distinguish three conceptually different sources of context.

Deployment-observable context. This is the intended setting of CAR: observations are grouped using information available independently of labels, such as temporal proximity, spatial adjacency, acquisition bursts, or externally defined episodes.

Analysis-time coherent context. Same-class and ordered contexts deliberately enforce strong coherence to test whether repeated observations can stabilize the adapter-response signal. These experiments answer a mechanism question, not a context-discovery question.

Reconstructed context from shuffled data. For the paired CIFAR diagnostic, shuffled batches are formed from a deterministic random permutation with seed $2026 + t$ at continual stage t, followed by consecutive groups of size 64. DBSCAN, feature kNN, adapter-signature kNN, and pseudo-class grouping then attempt to infer useful neighborhoods after destroying the original local ordering. Their weaker performance indicates that recovering semantic coherence from arbitrary shuffled observations is itself a difficult problem and should not be conflated with CAR's inference rule.

The distinction is important because CAR assumes that a useful local context is available; it does not claim to solve general-purpose unsupervised episode discovery.