CARE: LABEL-FREE COST-AWARE ONLINE ROUTING FOR ZERO-SHOT VLM CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive vision–language models (VLMs) exhibit complementary strengths and widely different computational costs in zero-shot classification, making a single fixed model suboptimal across diverse inputs. We introduce CARE, a label-free online framework for per-sample routing among heterogeneous CLIP-style VLMs under a global computational budget. Given a stream of unlabeled inputs, CARE dynamically selects one expert for each sample without requiring labeled calibration data or a pre-deployment reference set. To enable cross-model comparison, CARE standardizes logits on a per-sample basis and uses the resulting confidence as a label-free proxy reward. Routing is formulated as a contextual bandit with a dynamic dual update that balances estimated utility and computational cost while enforcing the global budget. Experiments across eight zero-shot classification benchmarks show that CARE achieves strong accuracy–cost trade-offs compared with individual models, full-expert methods, and routing baselines that use additional calibration or reference data.

1 INTRODUCTION

Contrastive VLMs, such as CLIP Radford et al. (2021), ALIGN Jia et al. (2021), and SigLIP Zhai et al. (2023), learn transferable representations from large-scale image–text pairs and enable strong zero-shot classification across diverse visual tasks. As the number of pretrained VLMs continues to grow through repositories such as Hugging Face Hugging Face, Inc. (2024) and OpenCLIP Ilharco et al. (2021), practitioners can choose from models with different architectures, scales, pre-training datasets, domain expertise, and computational costs. However, prior work has shown that VLM performance can vary substantially across tasks and domains, and that larger models are not uniformly superior to smaller or specialized alternatives Lu et al. (2023); Tan et al. (2025).

In practical zero-shot classification deployments, multiple VLMs may be available with different computational costs and complementary expertise. While larger or specialized models can provide higher accuracy on some inputs, they typically incur substantially higher inference cost, whereas smaller models may be sufficient for simpler instances. Selecting a single model offline can be reasonable when the deployment distribution is fixed and sufficient validation data are available, but such dataset-level selection cannot exploit variation in model suitability across individual samples and is less suitable for heterogeneous or changing input streams. Indeed, different VLMs can make correct predictions on different subsets of the same data Lu et al. (2023); Tan et al. (2025). These observations motivate per-sample routing that dynamically selects an appropriate expert for each input while allocating computation under a global budget.

A central challenge in label-free online VLM routing is that prediction correctness is unavailable at deployment, requiring the router to rely on confidence as a proxy for model utility. However, raw confidence scores are often not directly comparable across heterogeneous models. Even on the same input, different VLMs can produce logits with substantially different scales and sharpness due to differences in architecture, pre-training data, and adaptation strategy, causing some models to appear systematically more confident than others. Related scale mismatches have also been studied in knowledge distillation, where logit standardization has been used to remove differences in logit range and variance while preserving relative class information Sun et al. (2024).

To address these challenges, we propose CARE, a cost-aware, label-free framework for online per-sample routing among heterogeneous CLIP-style VLMs for zero-shot classification. CARE is built on two key ideas. First, because different VLMs can produce logits with substantially different scales and sharpness, we standardize logits on a per-sample basis to improve cross-model confidence comparability while preserving class ordering. Second, we use the resulting confidence scores as label-free proxy rewards within a contextual bandit formulation for sequential expert selection. A dynamic dual update balances estimated utility and computational cost while enforcing a global budget over the input stream. CARE therefore operates without target labels, labeled calibration data, or a pre-deployment reference set.

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.

We evaluate CARE on eight zero-shot classification benchmarks using a heterogeneous pool of general-purpose and domain-specialized CLIP-style VLMs spanning different architectures, pre-training data, and computational costs. Our evaluation includes comparisons with individual models, full-expert methods, online routing baselines, and routers that use additional reference or calibration data.

Our contributions are as follows:

- **Problem formulation.** We formulate online, label-free, cost-aware routing for zero-shot classification among heterogeneous CLIP-style VLMs as a sequential decision-making problem under a global budget constraint.
- **Cross-model confidence comparison.** We identify cross-model confidence comparability as a key challenge in label-free routing and show that per-sample logit standardization provides effective confidence-based utility signals without requiring calibration or reference data.
- **Budget-aware routing framework.** We propose CARE, a contextual bandit framework that uses confidence-derived proxy rewards and dynamic budget-aware decision making to adaptively route each input to a suitable expert.
- **Empirical validation.** Across eight zero-shot classification benchmarks, we show that CARE achieves strong accuracy–cost trade-offs compared with individual models, full-expert methods, and routing baselines, including methods that use additional reference or labeled calibration data.

2 Related work

Vision–language models. Recent years have seen rapid progress in pretrained vision–language models, including CLIP Radford et al. (2021), ALIGN Jia et al. (2021), and SigLIP Zhai et al. (2023). These models learn a shared embedding space between images and text, enabling strong zero-shot classification across diverse visual tasks. Building on these foundations, subsequent work has explored adapting VLMs through fine-tuning, prompt learning, and test-time methods Gao et al. (2024); Zhang et al. (2021); Zhou et al. (2022). At the same time, the VLM ecosystem has expanded to include both general-purpose and domain-specialized models for applications such as medical imaging and remote sensing Zhang et al. (2023); Liu et al. (2024). As a result, practitioners now have access to a diverse set of pretrained models with different architectures, scales, domain expertise, and computational costs, making effective model selection increasingly important.

Model selection, ensembling, and routing. Prior work on model selection has largely focused on offline and dataset-level settings. In the VLM literature, LOVM Zohar et al. (2023), Model Reuse Tan et al. (2025), and Ranked from Within Tu et al. (2025) aim to identify the best model for a target dataset using task descriptions, metadata, or label-free statistics. Beyond Sole Strength Lu et al. (2023) further shows that different VLMs often exhibit complementary strengths and can benefit from confidence-weighted ensembling. In the LLM literature, FrugalML Chen et al. (2020) employs a cascade of models, sequentially invoking increasingly expensive predictors until one produces a sufficiently confident output. MetaLLM Nguyen et al. (2024) and hybrid routing methods Ding et al. (2024) use learned routers or contextual bandits for model selection under cost constraints, while CaMVo Elumar et al. (2025) performs online cost-aware selection through subset voting. Recent work has also explored training-free nearest-neighbor routing Wu & Silwal (2025) and calibrated model cascades Zellinger & Thomson (2025). However, most existing approaches either operate offline, require pre-deployment reference or labeled calibration data, or rely on querying multiple models for each input.

Uncertainty estimation. Confidence and uncertainty signals are often used as proxies for model correctness in the absence of labels. In prior work on LLMs, Kadavath et al. (2022) uses self-consistency across multiple generated responses as a confidence signal, Li et al. (2024) uses token-level likelihoods to estimate answer reliability, and Azaria & Mitchell (2023) leverages internal hidden representations for uncertainty estimation and lie detection. In the VLM literature, Ranked from Within Tu et al. (2025) and Beyond Sole Strength Lu et al. (2023) show that confidence and uncertainty signals can be useful for ranking models or guiding ensemble weighting without labels. However, these signals are often difficult to compare across heterogeneous models because of differences in logit scale and sharpness. CARE uses per-sample logit standardization to improve cross-model confidence comparability without requiring labeled calibration or a pre-deployment reference set.

Table 1 summarizes the differences between CARE and prior model selection, ensembling, and routing approaches.

Table 1: Comparison with existing model selection and routing approaches.

Method	Domain	Pretrained	Granularity	Online	Cost-Aware	Target Dataset Information	Ensemble Type
LOVM Zohar et al. (2023)	VLM	Yes	Dataset-level	No	No	Dataset description	Single-model ranking
Model Reuse Tan et al. (2025)	VLM	Yes	Dataset-level	No	No	Dataset description	Top-k ensemble
Ranked from Within Tu et al. (2025)	VLM	No	Dataset-level	No	No	None	Single-model ranking
Beyond Sole Strength Lu et al. (2023)	VLM	No	Per-sample	No	No	None	Weighted ensemble
MetaLLM Nguyen et al. (2024)	LLM	Yes	Per-sample	Yes	Yes	Labeled validation data	Single-model routing
CaMVo Elumar et al. (2025)	LLM	No	Per-sample	Yes	Yes	None	Majority voting
CARE (ours)	VLM	No	Per-sample	Yes	Yes	None	Single-model routing

3 METHODOLOGY

3.1 ZERO-SHOT VISION–LANGUAGE MODELS

We consider zero-shot classification using CLIP-style VLMs, which consist of an image encoder and a text encoder that map both modalities into a shared embedding space. Let $\mathcal{I} : \mathcal{X} \to \mathbb{R}^d$ denote the image encoder and $\mathcal{T} : \mathcal{Y} \to \mathbb{R}^d$ denote the text encoder, where $\mathcal{X}$ is the image space, $\mathcal{Y}$ is the text space, and d is the dimension of the shared embedding space. Given an input image $x \in \mathcal{X}$ and a set of class names $\{y_j\}_{j=1}^C$, each class is converted into a text prompt $p_j \in \mathcal{Y}$ (e.g., "a photo of a $\{y_j\}$"). The model produces the image embedding $z^{\text{img}}(x) = \mathcal{I}(x)$ and the text embeddings $z_j^{\text{text}} = \mathcal{T}(p_j)$ for $j = 1, \ldots, C$. The model then computes cosine similarities between the image embedding and each text embedding as $\ell_j(x) = \langle z^{\text{img}}(x), z_j^{\text{text}} \rangle$ for $j = 1, \ldots, C$. Collecting these similarities across classes yields the logit vector $\ell(x) = \big(\ell_1(x), \ldots, \ell_C(x)\big) \in \mathbb{R}^C$. The predicted class is given by $\hat{y}(x) = \arg\max_{j \in [C]} \ell_j(x)$.

3.2 PROBLEM STATEMENT

We consider the problem of online routing among a heterogeneous pool of pretrained VLMs. Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ denote a collection of VLMs with different computational costs and performance characteristics. Each model m_k is associated with a normalized inference cost $c_{k,t} \in [0, 1]$, where larger models generally incur higher cost. Let $\{x_t\}_{t=1}^T$ denote a stream of unlabeled images arriving sequentially. At each round t, the learner observes an input image x_t and must select a single model $a_t \in [K]$ to produce a prediction. The selected model returns a logit vector $\ell_{a_t}(x_t)$ and incurs cost $c_{a_t,t}$. The routing decision must be made online, without access to ground-truth labels, calibration data, or domain annotations. Consequently, the learner cannot directly observe whether the selected model is correct and must instead rely on confidence signals derived from model outputs. Given a total budget $B > 0$, the objective is to maximize cumulative prediction quality while satisfying the resource constraint $\sum_{t=1}^T c_{a_t,t} \leq B$.

3.3 LOGIT STANDARDIZATION AND CONFIDENCE SCORES

Raw confidence scores are often difficult to compare reliably across heterogeneous VLMs because different models may produce logits with substantially different shifts and scales. These differences can arise from variations in architecture, pretraining data, embedding norms, and output sharpness. As a result, some models may appear systematically more confident simply because they produce logits with larger magnitude. To improve cross-model comparability, we standardize logits independently for each model and sample. Given a logit vector $\ell(x) \in \mathbb{R}^C$, we define the standardized logits as

$$\tilde{\ell}(x) = \frac{\ell(x) - \mu(\ell(x))\mathbf{1}}{\sigma(\ell(x))},$$

where $\mu(\ell(x))$ and $\sigma(\ell(x))$ denote the mean and standard deviation across classes, and $\mathbf{1} \in \mathbb{R}^C$ denotes the all-ones vector.

Proposition 1 (Properties of per-sample logit standardization)**.** *Let $T(\ell) = \frac{\ell - \mu(\ell)\mathbf{1}}{\sigma(\ell)}$ for $\sigma(\ell) > 0$. Then: (i) $T(a\mathbf{1} + b\ell) = T(\ell)$ for any $a \in \mathbb{R}$ and $b > 0$, (ii) $T(\ell)$ preserves class ordering, i.e., $\ell_i \leq \ell_j$ if and only if $T(\ell)_i \leq T(\ell)_j$, and (iii) $T(\ell)$ has zero mean and unit variance across classes.*

These properties make standardized logits particularly suitable for cross-model confidence comparison. Affine invariance removes model-specific shift and scale effects, rank preservation ensures that the predicted class and relative class ordering remain unchanged, and normalization maps all models into a common zero-mean, unit-variance space. Consequently, downstream confidence measures depend more on relative inter-class structure than on absolute logit magnitude. Importantly, the required statistics are computed from each individual sample and therefore require neither labeled calibration data nor a pre-deployment reference set.

Different confidence metrics can then be computed from the standardized logits $\tilde{\ell}(x)$. Let $\tilde{\ell}_{(1)}(x)$ denote the largest standardized logit and let $p_j(x) = \frac{\exp(\tilde{\ell}_j(x))}{\sum_{c=1}^{C}\exp(\tilde{\ell}_c(x))}$ denote the softmax probability of class j. We consider several confidence measures, including maximum logit $u_{\max}(x) = \tilde{\ell}_{(1)}(x)$, maximum softmax probability $u_{\text{softmax}}(x) = \max_j p_j(x)$, negative entropy $u_{\text{entropy}}(x) = \sum_{j=1}^{C} p_j(x)\log p_j(x)$, and energy $u_{\text{energy}}(x) = \log\sum_{j=1}^{C}\exp(\tilde{\ell}_j(x))$. In the routing framework described next, the resulting confidence score is used as a label-free proxy for model utility. In Section 4.3, we compare different confidence metrics and show that logit standardization substantially improves cross-model confidence comparison.

3.4 Budget-aware online routing

We propose CARE, a cost-aware online routing framework for zero-shot classification with heterogeneous VLMs. Given an input image, CARE selects a single VLM from the model pool by jointly considering its predicted reward and computational cost. To estimate model utility without access to labels, CARE uses confidence scores derived from standardized logits, which provide more reliable cross-model comparisons across heterogeneous VLMs. The routing decision is formulated as a contextual bandit problem, where the context is extracted from the input image and each arm corresponds to a candidate VLM. CARE maintains separate linear models for reward and cost estimation, and selects models according to an upper-confidence-bound criterion for reward and a lower-confidence-bound criterion for cost. These quantities are combined through a dynamic dual variable that enforces a global budget constraint. Intuitively, the dual variable increases when the algorithm overspends and decreases when the average cost remains below the target budget, allowing CARE to adaptively balance estimated utility and computational cost over time.

Contextual bandit formulation. We formalize CARE as a contextual bandit problem Li et al. (2010); Agrawal & Devanur (2016). At round t, the learner observes an input image x_t and computes a context vector $\phi_t = \mathcal{I}_{\text{base}}(x_t) \in \mathbb{R}^d$, where $\mathcal{I}_{\text{base}}$ denotes the image encoder of the base VLM. Selecting model m_k yields a confidence-based reward $r_{k,t}$ and incurs a normalized cost $c_{k,t} \in [0,1]$. In general, the cost may depend on both the selected model and the input image, for example due to differences in latency, image resolution or adaptive computation. Since only the selected model is evaluated, the learner observes $(r_{k,t}, c_{k,t})$ only for the chosen arm. We assume that the conditional mean reward and cost of each model depend linearly on the context. This choice is motivated by the high-dimensional feature representations produced by VLM encoders, for which linear contextual bandits provide a scalable framework. Formally,

$$\mathbb{E}[r_{k,t} \mid \phi_t] = \theta_k^{r\top}\phi_t, \qquad \mathbb{E}[c_{k,t} \mid \phi_t] = \theta_k^{c\top}\phi_t,$$

where $\theta_k^r, \theta_k^c \in \mathbb{R}^d$ are unknown model-specific parameters. The goal of the routing policy is to sequentially select models so that the cumulative reward $\sum_{t=1}^{T} r_{a_t,t}$ is maximized while ensuring that the total cost does not exceed the budget, i.e., $\sum_{t=1}^{T} c_{a_t,t} \le B$.

For each model m_k, CARE maintains separate linear estimators for reward and cost. Let $\lambda_0 > 0$ denote the ridge regularization parameter. For model m_k, we define $A_{k,t} = \lambda_0 I + \sum_{s<t:a_s=k}\phi_s\phi_s^\top$, $b_{k,t}^r = \sum_{s<t:a_s=k} r_{k,s}\phi_s$, $b_{k,t}^c = \sum_{s<t:a_s=k} c_{k,s}\phi_s$. The corresponding ridge regression estimates are given by $\hat{\theta}_{k,t}^r = A_{k,t}^{-1}b_{k,t}^r$, $\hat{\theta}_{k,t}^c = A_{k,t}^{-1}b_{k,t}^c$. In practice, the inverse matrices $A_{k,t}^{-1}$ are updated efficiently using the Sherman–Morrison formula Sherman & Morrison (1950), avoiding repeated matrix inversion at every round. Since each update has rank one,

$$A_{k,t+1}^{-1} = A_{k,t}^{-1} - \frac{A_{k,t}^{-1}\phi_t\phi_t^\top A_{k,t}^{-1}}{1+\phi_t^\top A_{k,t}^{-1}\phi_t}.$$

Therefore, updating $A_{k,t}^{-1}$ requires only $O(d^2)$ operations instead of recomputing the full inverse in $O(d^3)$ time.

Given the current context vector ϕ_t, CARE estimates an upper confidence bound on reward and a lower confidence bound on cost as

$$\text{UCB}_{k,t}^r = \hat{\theta}_{k,t}^{r\top}\phi_t + \alpha_r\sqrt{\phi_t^\top A_{k,t}^{-1}\phi_t}, \qquad \text{LCB}_{k,t}^c = \hat{\theta}_{k,t}^{c\top}\phi_t - \alpha_c\sqrt{\phi_t^\top A_{k,t}^{-1}\phi_t},$$

where $\alpha_r > 0$ and $\alpha_c > 0$ control the exploration strength for reward and cost, respectively. Intuitively, the reward estimate is optimistic while the cost estimate is pessimistic, encouraging CARE to favor models that are likely to achieve high confidence at low cost.

CARE combines the reward and cost estimates through a scalar dual variable $\lambda_t \in [0, \Lambda]$, where $\Lambda > 0$ denotes the maximum dual value. At round t, the score of model m_k is defined as $\text{Score}_{k,t} = \text{UCB}_{k,t}^r - \lambda_t\text{LCB}_{k,t}^c$. The selected

model is then given by $a_t = \arg\max_{k\in[K]} \mathrm{Score}_{k,t}$. Intuitively, the dual variable controls the trade-off between reward and cost. When λ_t is small, CARE places greater emphasis on reward and is more likely to select expensive high-performing models. As λ_t increases, the cost term is penalized more strongly, encouraging the selection of cheaper models. Let $\rho = B/T$ denote the target average cost per round. After selecting model a_t and observing $(r_{a_t,t}, c_{a_t,t})$, CARE updates the dual variable according to

$$\lambda_{t+1} = \Pi_{[0,\Lambda]}\left(\lambda_t + \eta(c_{a_t,t} - \rho)\right),$$

where $\eta > 0$ is the dual step size and $\Pi_{[0,\Lambda]}(\cdot)$ denotes projection onto the interval $[0, \Lambda]$. If selecting the chosen expert would exceed the remaining budget, CARE instead falls back to the base model, which incurs zero additional routing cost. The complete CARE procedure is summarized in Appendix C.

Theoretical discussion. CARE can be viewed as a single-resource specialization of contextual bandits with knapsacks. We follow the stochastic setting of LinCBwK Agrawal & Devanur (2016), assuming that the context and all experts' potential reward and cost vectors are drawn jointly i.i.d. across rounds from an unknown distribution. Let $\mathrm{OPT} := \mathbb{E}\left[\sum_{t=1}^{T} r_{\pi^\star(\phi_t),t}\right]$ denote the value of the optimal feasible static policy $\pi^\star$, which knows the joint distribution over contexts, rewards, and costs and selects a model based only on ϕ_t while satisfying the budget constraint in expectation over contexts, costs, and any policy randomization. We define the constrained regret as $\mathrm{Reg}(T) = \mathrm{OPT} - \sum_{t=1}^{T} r_{a_t,t}$.

Theorem 1 (Regret of CARE)**.** *Under the stochastic linear realizability and boundedness assumptions stated in Appendix B, with appropriate choices of the confidence parameters, dual step size, and dual cap, CARE satisfies, with probability at least* $1-\delta$*,*

$$\mathrm{Reg}(T) = \tilde{O}\left(\left(\frac{\mathrm{OPT}}{B} + 1\right) Kd\sqrt{T}\right),$$

where $\tilde{O}(\cdot)$ *hides logarithmic factors in* T*,* K*,* d*, and* $1/\delta$*.*

The theorem concerns an idealized sequential process that stops after a budget-reaching or budget-crossing action and assigns zero reward thereafter. The experimental implementation instead checks affordability before execution, permanently falls back to the base model with potentially nonzero reward, and uses mini-batched updates.

Theorem 1 adapts the LinCBwK analysis of Agrawal & Devanur (2016) to CARE's single-resource, arm-specific setting; the complete assumptions and proof are provided in Appendix B. Importantly, the regret is defined with respect to CARE's label-free confidence-based proxy reward, rather than ground-truth classification accuracy. Thus, the theorem characterizes how closely CARE approaches the optimal feasible routing policy under this observable proxy objective. In Section 4.3, we empirically evaluate the relationship between the proxy reward and prediction correctness, showing that higher confidence is generally associated with higher classification accuracy.

4 Experiments

4.1 Experimental setup

Datasets and model zoo. We evaluate CARE on eight datasets spanning natural images, remote sensing, traffic signs, texture recognition, fine-grained categorization, and computational pathology: CIFAR-10 Krizhevsky (2009), RESISC45 Cheng et al. (2017), PCam Veeling et al. (2018), GTSRB Stallkamp et al. (2011), Food101 Bossard et al. (2014), DTD Cimpoi et al. (2014), Flowers102 Nilsback & Zisserman (2008), and CRC Kather et al. (2019). We use a heterogeneous model zoo consisting of general-purpose CLIP models, high-capacity vision transformers, and domain-specialized experts.

Baselines. We organize the compared methods according to their inference and pre-deployment information requirements. **Individual models** evaluate each VLM independently on all inputs. **Full-expert methods** include a Confidence Oracle, which selects the highest-confidence expert after evaluating the full model pool, and a confidence-weighted ensemble inspired by Lu et al. (2023). **Routing without pre-deployment data** includes the Fixed Cascade, Neural-LinUCB Xu et al. (2022), and CARE; these methods require neither a separate reference set nor labeled calibration data. Finally, **routing with pre-deployment data** includes a PORT-inspired nearest-neighbor router Wu & Silwal (2025), which uses an unlabeled reference set with offline expert profiling, and a supervised calibrated cascade inspired by Zellinger & Thomson (2025), which uses labeled data for confidence calibration and threshold selection. Detailed baseline definitions and implementation settings are provided in Appendix D.5.

Implementation details. CARE operates in a fully online and label-free setting. For each sample, the base model is evaluated to obtain contextual features and confidence scores, and the router selects one additional expert under a global budget $B = \rho T$. In our experiments, expert costs are fixed log-normalized FLOP-based quantities known before routing, and we use normalized entropy as the default proxy reward. For computational efficiency, routing parameters are updated in mini-batches of size 128. Unless otherwise stated, we use $\alpha_r = 1.0$, $\eta = 0.5$, and $\lambda_{\max} = 5.0$. Sensitivity to α_r, η, and batch size is reported in Appendix G, while runtime overhead and per-model inference latency are reported in Appendix I. All primary results use a fixed test stream, with robustness to input ordering evaluated separately in Appendix F. Additional implementation details are provided in Appendix D.

4.2 RESULTS

Table 2 compares CARE with individual models, full-expert methods, online routing baselines, and routers that use additional pre-deployment data. CARE yields particularly large gains over the base model on datasets where specialist experts are available. On CRC, CARE-60 improves accuracy from 11.66% to 35.54%, a gain of nearly 24 percentage points, by selectively routing samples mostly to BiomedCLIP. Similarly, on PCam, CARE-60 improves over the base model from 58.79% to 66.55%, while on RESISC45 it increases accuracy from 55.25% to 68.65% by exploiting RemoteCLIP. Flowers102 also benefits substantially from routing, with CARE-60 reaching 81.14% compared with 64.81% for the base model.

Directly selecting a single fixed VLM is difficult in practice because the strongest model varies substantially across datasets, while CARE operates without labeled validation data for model selection. Moreover, uniformly deploying the strongest model would ignore the explicit computational budget. CARE instead adapts online to the incoming samples and available budget. For example, on CIFAR-10, CARE-60 reaches 95.24% accuracy at a cost of 0.59, compared with 98.62% for CLIP ViT-H/14-378 at cost 1.0. On Food101, CARE-100 reaches 95.46%, close to the 96.19% accuracy of ViT-H/14-378, while on Flowers102 it achieves 87.77% compared with 92.15% for the largest expert. We also evaluate CARE on a mixed-domain stream without providing domain identity; detailed results are reported in Appendix H.

Compared with other methods that require no separate pre-deployment dataset, CARE provides a strong accuracy–cost trade-off. CARE-60 improves average accuracy over the Fixed Cascade from 64.48% to 67.36% at essentially the same average cost (0.58), while CARE-100 reaches 69.92% average accuracy. Neural-LinUCB achieves 67.47% at a cost of 0.68, indicating that introducing a nonlinear contextual reward representation alone does not improve over CARE's lightweight linear formulation.

The full-expert baselines achieve strong accuracy but require evaluating the entire model pool for every input, resulting in an average additional cost of 3.34. CARE-100 achieves 69.92% average accuracy at a cost of 0.72 while recovering much of their predictive performance. For example, on CIFAR-10 it achieves 95.69% compared with 95.71% for the Confidence Oracle, and on GTSRB it reaches 51.23% compared with 51.40%, while requiring substantially less computation.

Finally, CARE remains competitive with routing methods that receive additional pre-deployment information. CARE-100 achieves 69.92% average accuracy at 0.72 cost, compared with 69.06% at 0.74 for the PORT-inspired router, despite requiring no unlabeled reference bank or offline full-expert profiling. The supervised calibrated cascade achieves 70.22% at 0.70 cost using labeled calibration data. Thus, CARE approaches the performance of a calibrated routing policy while retaining a fully label-free online deployment setting.

4.3 ABLATION STUDIES

Confidence metrics. Confidence metric choice has a noticeable impact on both oracle selection and routing performance, especially when models with different logit scales are combined. We compare maximum logit, maximum softmax probability, negative entropy, and energy, using both raw and per-sample standardized logits. Each metric is used as the label-free routing reward. Table 3 reports the average accuracy across all eight datasets. Standardization substantially improves both oracle and routing performance for most metrics, highlighting the importance of cross-model logit comparability. Among the standardized variants, energy achieves the highest oracle accuracy, while maximum logit achieves the highest CARE accuracy. Detailed per-dataset results are provided in Appendix E.

Confidence as a Proxy for Correctness. CARE relies on confidence as a label-free proxy for prediction quality. To evaluate this relationship, for each sample we select the expert with the highest standardized confidence and record both its confidence and whether its prediction is correct. We compute the AUROC between confidence and correctness and additionally partition samples into five confidence quintiles. Table 4 shows a meaningful, though imperfect, relationship between confidence and correctness, with an average AUROC of 0.7508. Average accuracy also generally increases

Method	CIFAR-10		RESISC45		PCam		GTSRB		Food101		DTD		Flowers102		CRC		Avg.	
	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost
Individual models																		
Base (ViT-B/32)	88.79	0.00	55.25	0.00	58.79	0.00	28.09	0.00	80.19	0.00	42.02	0.00	64.81	0.00	11.66	0.00	53.70	0.00
RemoteCLIP	37.61	0.39	68.46	0.39	52.83	0.39	5.73	0.39	35.37	0.39	20.37	0.39	29.47	0.39	24.78	0.39	34.33	0.39
BiomedCLIP	23.75	0.53	16.19	0.53	75.68	0.53	2.00	0.53	1.92	0.53	10.00	0.53	0.89	0.53	42.70	0.53	21.64	0.53
ConvNeXt-Large-D	96.46	0.68	69.60	0.68	54.46	0.68	44.66	0.68	90.59	0.68	64.47	0.68	78.21	0.68	31.04	0.68	66.19	0.68
CLIP ViT-L/14	95.32	0.74	65.92	0.74	55.47	0.74	48.65	0.74	89.79	0.74	51.91	0.74	76.86	0.74	50.21	0.74	66.77	0.74
CLIP ViT-H/14-378	98.62	1.00	75.56	1.00	56.88	1.00	60.04	1.00	96.19	1.00	66.70	1.00	92.15	1.00	48.15	1.00	74.29	1.00
Full-expert methods																		
Confidence Oracle	95.71	3.34	75.24	3.34	75.96	3.34	51.40	3.34	95.26	3.34	64.15	3.34	89.93	3.34	40.92	3.34	73.57	3.34
Weighted Ensemble	97.19	3.34	79.03	3.34	65.50	3.34	51.66	3.34	92.68	3.34	63.19	3.34	81.98	3.34	44.93	3.34	72.02	3.34
Routing without pre-deployment data																		
Fixed Cascade	94.66	0.71	70.14	0.59	64.75	0.62	41.22	0.58	91.28	0.72	51.96	0.65	67.44	0.25	34.37	0.54	64.48	0.58
Neural-LinUCB	95.45	0.68	64.05	0.60	62.44	0.53	49.23	0.66	92.21	0.90	54.19	0.67	85.21	0.83	37.02	0.57	67.47	0.68
Ours (CARE-20)	91.50	0.20	60.57	0.20	62.83	0.20	31.22	0.20	83.99	0.20	41.60	0.20	67.80	0.20	22.49	0.20	57.75	0.20
Ours (CARE-60)	95.24	0.59	68.65	0.58	66.55	0.56	46.39	0.60	91.25	0.60	54.10	0.60	81.14	0.60	35.54	0.51	67.36	0.58
Ours (CARE-100)	95.69	0.70	68.75	0.65	66.26	0.56	51.23	0.73	95.46	0.96	57.07	0.73	87.77	0.93	37.09	0.52	69.92	0.72
Routing with pre-deployment data																		
PORT-inspired kNN	95.73	0.68	67.71	0.76	64.00	0.66	49.14	0.69	95.31	0.95	56.12	0.77	87.84	0.75	36.62	0.62	69.06	0.74
Calibrated Cascade	95.87	0.75	70.63	0.68	67.08	0.58	50.78	0.76	95.11	0.80	57.02	0.73	88.01	0.70	37.23	0.59	70.22	0.70

Table 2: Zero-shot classification accuracy (%) and log-normalized additional routing cost. Methods are grouped according to their inference and pre-deployment information requirements. CARE-x denotes CARE with target normalized routing-cost level $\rho = x/100$.

Metric	Oracle		CARE	
	Raw	Std.	Raw	Std.
Max Logit	24.75	73.54	25.78	**67.56**
Max Softmax	72.14	73.42	55.76	67.46
Neg. Entropy	51.73	73.60	39.45	67.36
Energy	22.36	**73.70**	29.03	67.11

Table 3: Average oracle and CARE accuracy across eight datasets using raw and standardized confidence metrics. CARE uses $\rho = 0.6$.

from 55.01% in the lowest-confidence quintile to 90.32% in the highest. These findings indicate that standardized confidence provides useful information about the correctness of the selected predictions, supporting its use as a proxy for prediction quality in CARE.

Statistic	CIFAR-10	RESISC45	PCam	GTSRB	Food101	DTD	Flowers102	CRC	Avg.
AUROC	0.7902	0.7505	0.7021	0.7751	0.8091	0.7633	0.8063	0.6101	0.7508
0–20%	87.05	50.58	55.70	25.18	85.62	34.04	68.05	33.84	55.01
20–40%	96.02	62.25	69.90	28.98	94.95	52.93	92.26	28.83	65.77
40–60%	98.00	78.89	77.40	46.67	98.05	61.15	95.77	36.21	74.02
60–80%	98.55	88.57	84.45	69.48	98.71	77.39	95.85	49.03	82.75
80–100%	98.95	95.87	92.40	86.70	99.01	95.21	97.72	56.69	90.32

Table 4: Confidence–correctness relationship across datasets.

Accuracy-cost tradeoff. Figure 1 illustrates the accuracy–cost trade-off of CARE across routing budgets with $\rho \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$. In most datasets, the CARE operating points form a smooth Pareto frontier, indicating that increasing the routing budget consistently improves accuracy without introducing dominated operating points. This behavior highlights that CARE can adaptively trade computation for performance and recover much of the benefit of stronger experts while remaining substantially cheaper than the Confidence Oracle.

Expert utilization analysis. Figure 2 shows that CARE learns dataset-specific routing patterns as the available budget increases. At low budgets, most samples are assigned to the base model, while larger budgets gradually shift usage toward stronger experts and larger generalist VLMs. For general-domain datasets such as CIFAR-10, Food101, and Flowers102, CARE increasingly relies on high-capacity generalist models, particularly CLIP ViT-H/14-378 and CLIP

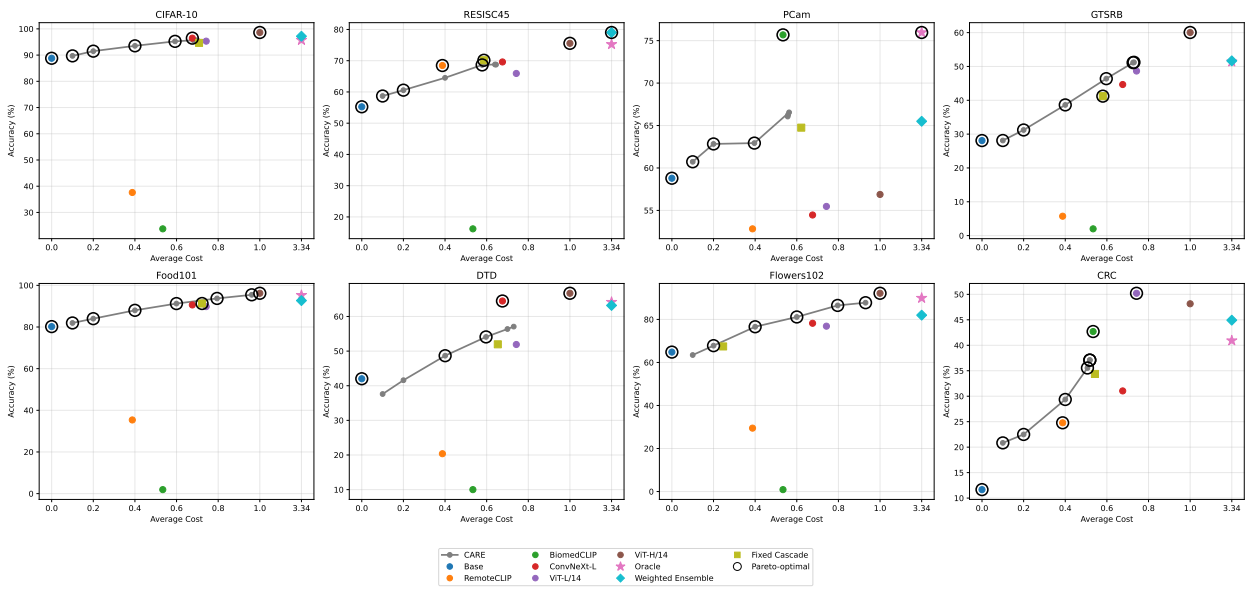

Figure 1: Accuracy–cost tradeoff across datasets. Pareto-optimal operating points are highlighted with black circles.

ViT-L/14, as more budget becomes available. In contrast, specialist experts dominate on domain-specific datasets. BiomedCLIP receives a large fraction of the routing budget on PCam and CRC, while RemoteCLIP becomes the dominant specialist on RESISC45. ConvNeXt-Large-D is also frequently selected on DTD and GTSRB, where texture and local visual patterns are especially important. These results suggest that CARE successfully identifies the most suitable expert for each domain and allocates additional computation accordingly.

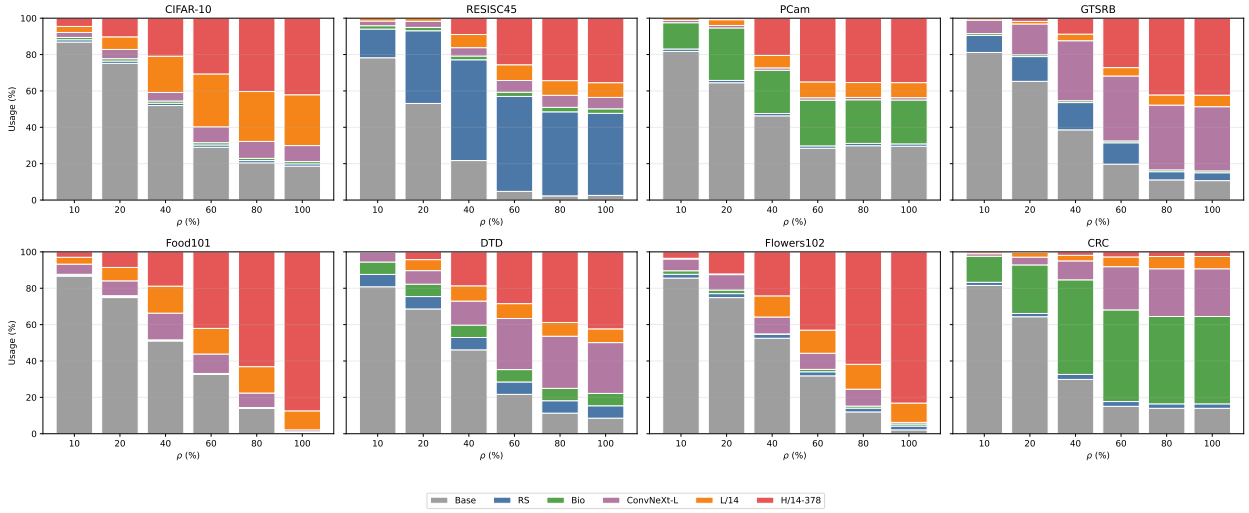

Figure 2: Model utilization across routing budgets. Each subplot shows the fraction of samples assigned to each expert as the average cost budget ρ increases. CARE adapts its routing policy across datasets, gradually shifting from the low-cost base model toward stronger specialists and larger generalist VLMs as more budget becomes available.

Accuracy and cost over batches. Figure 3 illustrates the cumulative routing behavior of CARE-60 over time. At the beginning of routing, cumulative accuracy is often unstable due to initial exploration and limited observations for estimating expert rewards. As more batches are processed, the routing policy gradually identifies stronger experts, leading to improved and more stable cumulative accuracy. The cumulative average cost also remains consistently below the target threshold of $\rho = 0.6$ across datasets, indicating that the dynamic dual variable effectively controls budget

usage during online routing. In most cases, the average cost quickly approaches the desired budget level and then stabilizes, showing that CARE can maintain strong performance while respecting its cost constraint throughout the entire routing process. Additional results for other budget levels are provided in Appendix J.

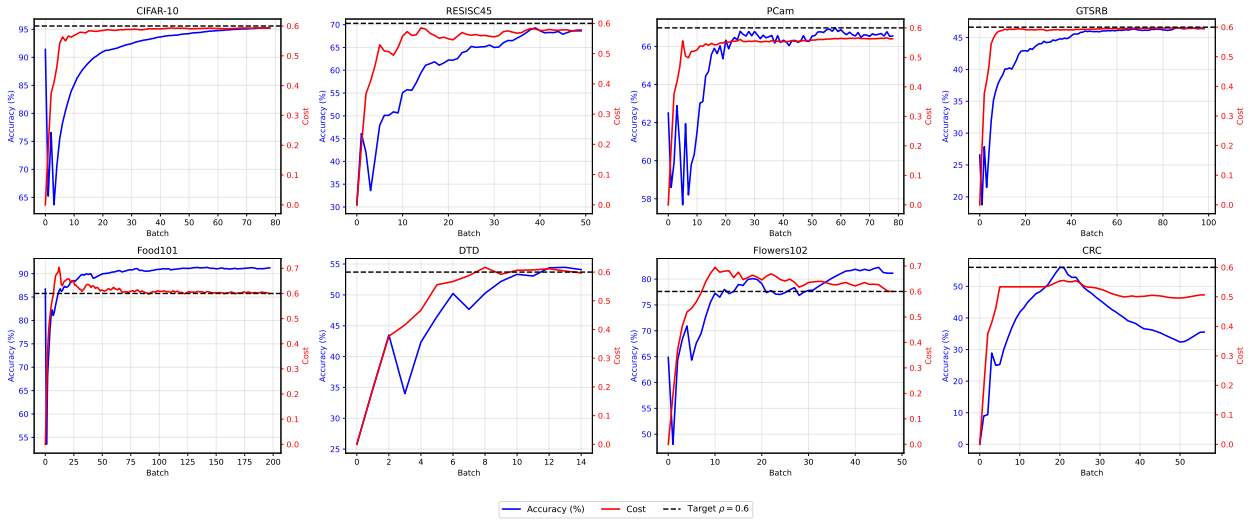

Figure 3: Cumulative accuracy and average cost over routing rounds for CARE-60. In each subplot, the blue curve shows cumulative accuracy and the red curve shows cumulative average cost. The dashed black line indicates the target budget level $\rho = 0.6$.

5 Conclusion and Limitations

In this paper, we introduced CARE, a label-free and budget-aware framework for online per-sample routing among heterogeneous CLIP-style VLMs for zero-shot classification. CARE combines per-sample logit standardization with a contextual bandit formulation to estimate model utility from confidence and select experts under a global routing-cost budget. Across eight zero-shot classification benchmarks, CARE achieves strong accuracy–cost trade-offs without requiring labeled calibration data or a pre-deployment reference set. Our ablation studies show that logit standardization improves performance on average and that standardized confidence is positively associated with prediction correctness. Expert-utilization patterns further illustrate how CARE's model selection varies across datasets and routing budgets.

Limitations. CARE relies on confidence as a label-free proxy for model utility rather than directly observing classification correctness. This is inherent to our label-free setting, where ground-truth feedback is unavailable during deployment. Although our empirical analysis shows a meaningful relationship between confidence and correctness, highly confident errors can still occur, motivating improved label-free uncertainty estimators.

CARE also requires evaluating a shared base VLM for every input to obtain the routing context. This computation enables per-sample routing without evaluating the full expert pool, and our experiments show favorable accuracy–cost trade-offs despite this overhead. Nevertheless, the shared base-model cost may be relevant in highly resource-constrained deployments.

Finally, our regret analysis relies on stochastic linear realizability and boundedness assumptions commonly used in contextual bandit analysis. The empirical results suggest that the resulting linear estimator is effective in our evaluation setting, but real-world streams and reward relationships may deviate from these assumptions. Extending the analysis to more general reward and cost models is therefore an interesting direction for future work.

AI use statement

We used generative AI tools to assist with formulating mathematical claims, developing key ingredients of proofs, writing proofs, proposing or refining hypotheses, providing feedback on research methodology and experiments, implementing methods, and interpreting results. We did not use generative AI tools to generate synthetic datasets or develop the underlying theoretical model or conceptual framework. Translation, dataset cleaning and reformatting, and

qualitative or thematic data analysis were not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of the manuscript, summarize or analyze existing literature, discover research topics or identify gaps, and identify relevant literature. The authors reviewed all AI-assisted work, including checking the code, mathematical claims and proofs, references, and manuscript text. We take responsibility for the final content of this work, including all text, mathematical claims, proofs, and code produced with the assistance of generative AI.

Reproducibility statement

CARE's routing procedure is provided in Appendix C (Algorithm 1). Section 4.1 describes the experimental setup and main hyperparameter settings, while Appendix D provides additional details on the model pool, prompt templates, cost and reward construction, baseline implementations, and computing environment. The assumptions and proofs supporting the regret analysis are provided in Appendix B. Additional evaluations of sensitivity to hyperparameters and test-stream ordering are reported in Appendices G and F, respectively. We also submit our implementation as supplementary material to support reproduction of the experimental results.

References

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

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

Lingjiao Chen, Matei Zaharia, and James Y Zou. Frugalml: How to use ml prediction apis more accurately and cheaply. *Advances in neural information processing systems*, 33:10685–10696, 2020.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024.

Eray Can Elumar, Cem Tekin, and Osman Yagan. Cost-aware llm-based online dataset annotation. *arXiv preprint arXiv:2505.15101*, 2025.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2): 581–595, 2024.

Hugging Face, Inc. Hugging face hub. `https://huggingface.co`, 2024.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Christian-Alexander Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dominik Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):e1002730, 2019.

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

Jia Li, Yuqi Zhu, Yongmin Li, Ge Li, and Zhi Jin. Showing llm-generated code selectively based on confidence of llms. *arXiv preprint arXiv:2410.03234*, 2024.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

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

Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. *arXiv preprint arXiv:2311.17091*, 2023.

Quang H Nguyen, Thinh Dao, Duy C Hoang, Juliette Decugis, Saurav Manchanda, Nitesh V Chawla, and Khoa D Doan. Metallm: A high-performant and cost-efficient dynamic framework for wrapping llms. *arXiv preprint arXiv:2407.10834*, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

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

Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740, 2024. doi: 10.1109/CVPR52733.2024.01489.

Hao-Zhe Tan, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Vision-language model selection and reuse for downstream adaptation. *arXiv preprint arXiv:2501.18271*, 2025.

Weijie Tu, Weijian Deng, Dylan Campbell, Yu Yao, Jiyang Zheng, Tom Gedeon, and Tongliang Liu. Ranked from within: Ranking large multimodal models without labels. In *International Conference on Machine Learning*, pp. 60362–60384. PMLR, 2025.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. *arXiv preprint arXiv:1806.03962*, 2018.

Fangzhou Wu and Sandeep Silwal. Efficient training-free online routing for high-volume multi-llm serving. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4577.

Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration. In *International Conference on Learning Representations*, 2022.

Michael J. Zellinger and Matt Thomson. Rational tuning of llm cascades via probabilistic modeling. *Transactions on Machine Learning Research*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022.

Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. Lovm: Language-only vision model selection. *Advances in Neural Information Processing Systems*, 36:33120–33132, 2023.

A Proof of proposition 1

Proof. Let $\tilde{\ell} = a\mathbf{1} + b\ell$ with $b > 0$. Then

$$\mu(\tilde{\ell}) = a + b\mu(\ell), \qquad \sigma(\tilde{\ell}) = b\sigma(\ell).$$

Therefore,

$$T(\tilde{\ell}) = \frac{\tilde{\ell} - \mu(\tilde{\ell})\mathbf{1}}{\sigma(\tilde{\ell})} = \frac{a\mathbf{1} + b\ell - (a + b\mu(\ell))\mathbf{1}}{b\sigma(\ell)} = \frac{\ell - \mu(\ell)\mathbf{1}}{\sigma(\ell)} = T(\ell),$$

which proves affine invariance.

Since $T(\ell)$ is obtained by subtracting the same scalar from every coordinate and dividing by a positive scalar, class ordering is preserved:

$$\ell_i \le \ell_j \quad \Longleftrightarrow \quad T(\ell)_i \le T(\ell)_j.$$

Finally, by construction,

$$\frac{1}{C}\sum_{j=1}^{C} T(\ell)_j = 0, \qquad \frac{1}{C}\sum_{j=1}^{C} (T(\ell)_j)^2 = 1.$$

Thus, standardized logits have zero mean and unit variance. □

B Regret analysis

In this section, we provide a regret guarantee for the sequential stochastic formulation of CARE. Our analysis follows the proof template of LinCBwK Agrawal & Devanur (2016) and is primarily an adaptation rather than a new bandit-theoretic contribution. The main structural difference is that CARE operates in a single-resource setting, which allows the vector-valued dual variable in LinCBwK to be replaced by a scalar dual variable with a projected update.

At the proof level, this leads to three concrete differences. First, the dual update is analyzed via a one-dimensional projected online gradient descent argument instead of the multiplicative-weights analysis used in LinCBwK. Second, the surrogate comparison simplifies due to the scalar dual. Third, CARE uses disjoint arm-specific linear models, which can be embedded into a shared linear representation of dimension Kd.

Assumptions. Let H_{t-1} denote the history available before observing the context at round t. We assume that, for all rounds t and experts m_k,

$$\|\phi_t\|_2 \le 1, \qquad r_{k,t} \in [0,1], \qquad c_{k,t} \in [0,1].$$

The conditional mean reward and cost are linear in the observed context:

$$\mu^r_{k,t} := \mathbb{E}[r_{k,t} \mid H_{t-1}, \phi_t] = \theta^{r\top}_k \phi_t,$$

$$\mu^c_{k,t} := \mathbb{E}[c_{k,t} \mid H_{t-1}, \phi_t] = \theta^{c\top}_k \phi_t,$$

where

$$\|\theta^r_k\|_2 \le 1, \qquad \|\theta^c_k\|_2 \le 1.$$

Define the post-context filtration

$$\mathcal{G}_t := \sigma(H_{t-1}, \phi_t).$$

Since the routing action a_t is selected after observing ϕ_t, it is $\mathcal{G}_t$-measurable. Consequently,

$$\mathbb{E}\left[r_{a_t,t} - \mu^r_{a_t,t} \mid \mathcal{G}_t\right] = 0,$$

and

$$\mathbb{E}\left[c_{a_t,t} - \mu^c_{a_t,t} \mid \mathcal{G}_t\right] = 0.$$

We assume that, across rounds,

$$\{(\phi_t, r_{k,t}, c_{k,t})\}_{k=1}^{K}$$

is generated independently from an unknown stationary joint distribution $\mathcal{D}$. Since rewards and costs are bounded, the corresponding centered noise variables are conditionally sub-Gaussian.

Let

$$\rho := \frac{B}{T}, \qquad 0 < \rho \leq 1.$$

For the theoretical analysis, we additionally introduce a no-op action $k = 0$ satisfying

$$r_{0,t} = c_{0,t} = 0, \qquad \theta_0^r = \theta_0^c = 0.$$

The no-op action represents declining to invoke an additional expert and ensures that a feasible zero-cost action is always available. We define its reward and cost surrogates to be zero.

Stopping convention. For the sequential theoretical process, define

$$\tau = \min\left\{ t \leq T : \sum_{s=1}^{t} c_{a_s,s} \geq B \right\} \wedge T.$$

After round τ, the theoretical process selects the no-op action for all remaining rounds.

Define

$$I_t := \mathbf{1}\{t \leq \tau\}.$$

Because whether round t is active depends only on costs observed before round t, I_t is H_{t-1}-measurable. For analytical purposes, the projected dual update is also defined after the terminal round τ so that $\lambda_{\tau+1}$ is well defined, although it is not used for any subsequent routing decision.

Benchmark policy. As the competitor class, we consider context-dependent static policies. A policy π maps each context ϕ to a probability distribution over $\{0, 1, \ldots, K\}$, where $\pi_k(\phi)$ denotes the probability assigned to action k.

For horizon T and budget B, let

$$\pi^\star = \arg\max_{\pi} T\mathbb{E}_{\mathcal{D}}\left[\sum_{k=0}^{K} \pi_k(\phi)\theta_k^{r\top}\phi\right]$$

subject to

$$T\mathbb{E}_{\mathcal{D}}\left[\sum_{k=0}^{K} \pi_k(\phi)\theta_k^{c\top}\phi\right] \leq B.$$

Let OPT denote the value of $\pi^\star$. Define

$$\bar{r}^\star := \frac{\mathrm{OPT}}{T},$$

and

$$\bar{c}^\star := \mathbb{E}_{\mathcal{D}}\left[\sum_{k=0}^{K} \pi_k^\star(\phi)\theta_k^{c\top}\phi\right] \leq \rho.$$

We define the constrained regret as

$$\mathrm{Reg}(T) = \mathrm{OPT} - \sum_{t=1}^{T} r_{a_t,t}.$$

Disjoint-arm reduction. For each expert k, define the block feature

$$\widetilde{\phi}_{k,t} = (0, \ldots, 0, \phi_t, 0, \ldots, 0) \in \mathbb{R}^{Kd},$$

where ϕ_t occupies the k-th block. Concatenating the arm-specific parameters gives

$$\Theta^r = [\theta_1^{r\top}, \ldots, \theta_K^{r\top}]^\top, \qquad \Theta^c = [\theta_1^{c\top}, \ldots, \theta_K^{c\top}]^\top,$$

so that

$$\Theta^{r\top}\widetilde{\phi}_{k,t} = \theta_k^{r\top}\phi_t, \qquad \Theta^{c\top}\widetilde{\phi}_{k,t} = \theta_k^{c\top}\phi_t.$$

We use the conservative effective dimension

$$D = Kd.$$

For each expert, define

$$s_{k,t} := \sqrt{\phi_t^\top A_{k,t}^{-1}\phi_t}.$$

CARE constructs the optimistic reward and pessimistic cost surrogates

$$\widetilde{r}_{k,t} := \widehat{\theta}_k^{r\top}\phi_t + \alpha_r s_{k,t},$$

and

$$\widetilde{c}_{k,t} := \widehat{\theta}_k^{c\top}\phi_t - \alpha_c s_{k,t}.$$

These confidence surrogates need not themselves lie in $[0,1]$.

Lemma 1 (Projected dual-update regret)**.** *Let*

$$\lambda_{t+1} = \Pi_{[0,\Lambda]}\left(\lambda_t + \eta z_t\right), \qquad z_t \in [-1,1].$$

Then, for any $\lambda \in [0,\Lambda]$ and any $\tau \le T$,

$$\sum_{t=1}^{\tau}\lambda z_t - \sum_{t=1}^{\tau}\lambda_t z_t \le \frac{\Lambda^2}{2\eta} + \frac{\eta T}{2}.$$

Proof. Since Euclidean projection onto a convex set is non-expansive,

$$|\lambda_{t+1} - \lambda|^2 \le |\lambda_t + \eta z_t - \lambda|^2.$$

Expanding the right-hand side,

$$|\lambda_{t+1} - \lambda|^2 \le |\lambda_t - \lambda|^2 + 2\eta z_t(\lambda_t - \lambda) + \eta^2 z_t^2.$$

Rearranging,

$$z_t(\lambda - \lambda_t) \le \frac{|\lambda_t - \lambda|^2 - |\lambda_{t+1} - \lambda|^2}{2\eta} + \frac{\eta}{2}z_t^2.$$

Summing over $t = 1,\dots,\tau$, dropping the final non-positive term, and using $|\lambda_1 - \lambda| \le \Lambda$ and $z_t^2 \le 1$ gives

$$\sum_{t=1}^{\tau}\lambda z_t - \sum_{t=1}^{\tau}\lambda_t z_t \le \frac{\Lambda^2}{2\eta} + \frac{\eta T}{2}.$$

□

Lemma 2 (Actual-cost dual control)**.** *Suppose the dual variable is updated using the realized cost,*

$$\lambda_{t+1} = \Pi_{[0,\Lambda]}\left(\lambda_t + \eta(c_{a_t,t} - \rho)\right).$$

Then

$$\sum_{t=1}^{\tau}\lambda_t(c_{a_t,t} - \rho) \ge \Lambda(B - \tau\rho) - \left(\frac{\Lambda^2}{2\eta} + \frac{\eta T}{2}\right).$$

Proof. Apply Lemma 1 with

$$z_t = c_{a_t,t} - \rho.$$

Since $c_{a_t,t}, \rho \in [0,1]$,

$$z_t \in [-1,1].$$

Thus, for any $\lambda \in [0,\Lambda]$,

$$\sum_{t=1}^{\tau}\lambda_t(c_{a_t,t} - \rho) \ge \lambda\sum_{t=1}^{\tau}(c_{a_t,t} - \rho) - \left(\frac{\Lambda^2}{2\eta} + \frac{\eta T}{2}\right).$$

If $\tau < T$, choose $\lambda = \Lambda$. By the definition of τ,

$$\sum_{t=1}^{\tau} c_{a_t,t} \geq B,$$

and therefore

$$\lambda \sum_{t=1}^{\tau} (c_{a_t,t} - \rho) \geq \Lambda(B - \tau\rho).$$

If $\tau = T$, choose $\lambda = 0$. Since

$$B - \tau\rho = B - T(B/T) = 0,$$

the same bound holds.

Combining the two cases proves the result. □

Confidence and concentration event. For each arm-specific d-dimensional ridge regression, standard self-normalized concentration implies that, for suitably chosen α_r and α_c, simultaneously for all rounds and experts,

$$\left| \widehat{\theta}_k^{r\top} \phi_t - \mu_{k,t}^r \right| \leq \alpha_r s_{k,t},$$

and

$$\left| \widehat{\theta}_k^{c\top} \phi_t - \mu_{k,t}^c \right| \leq \alpha_c s_{k,t}.$$

For example, up to universal constants it is sufficient to take

$$\alpha_r, \alpha_c = O\left(\sqrt{d \log(1 + T/\gamma) + \log(K/\delta)} + \sqrt{\gamma} \right).$$

Consequently,

$$0 \leq \widetilde{r}_{k,t} - \mu_{k,t}^r \leq 2\alpha_r s_{k,t},$$

and

$$0 \leq \mu_{k,t}^c - \widetilde{c}_{k,t} \leq 2\alpha_c s_{k,t}.$$

Define

$$R_r(T) := 2\alpha_r \sum_{t=1}^{\tau} s_{a_t,t},$$

and

$$R_c(T) := 2\alpha_c \sum_{t=1}^{\tau} s_{a_t,t}.$$

Using the elliptical-potential lemma for the disjoint arm-specific regressions gives the conservative bounds

$$R_r(T), R_c(T) = \widetilde{O}(Kd\sqrt{T}).$$

We will also use concentration for several bounded martingale-difference sequences. Let

$$M_T(\delta) = O\left(\sqrt{T \log(1/\delta)} \right).$$

Each confidence or martingale event below is constructed with failure probability at most δ/C for a fixed universal constant C; a union bound therefore gives a joint event of probability at least $1 - \delta$.

On this event,

$$\left| \sum_{t=1}^{T} I_t (r_{a_t,t} - \mu_{a_t,t}^r) \right| \leq M_T(\delta),$$

and

$$\left| \sum_{t=1}^{T} I_t \lambda_t (c_{a_t,t} - \mu_{a_t,t}^c) \right| \leq \Lambda M_T(\delta).$$

For the benchmark policy, define

$$R_t^\star = \sum_{k=0}^{K} \pi_k^\star(\phi_t)\mu_{k,t}^r,$$

and

$$C_t^\star = \sum_{k=0}^{K} \pi_k^\star(\phi_t)\mu_{k,t}^c.$$

By the i.i.d. assumption,

$$\mathbb{E}[R_t^\star \mid H_{t-1}] = \bar{r}^\star = \frac{\mathrm{OPT}}{T},$$

and

$$\mathbb{E}[C_t^\star \mid H_{t-1}] = \bar{c}^\star \le \rho.$$

Since I_t and λ_t are H_{t-1}-measurable, concentration of the corresponding predictable martingale sums gives

$$\sum_{t=1}^{T} I_t R_t^\star \ge \frac{\tau}{T}\mathrm{OPT} - M_T(\delta),$$

and

$$\sum_{t=1}^{T} I_t \lambda_t (C_t^\star - \rho) \le \Lambda M_T(\delta).$$

Lemma 3 (Surrogate comparison)**.** *On the joint confidence and concentration event,*

$$\sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} \ge \frac{\tau}{T}\mathrm{OPT} + \sum_{t=1}^{\tau} \lambda_t(\widetilde{c}_{a_t,t} - \rho) - (1+\Lambda)M_T(\delta).$$

Proof. At each active round, CARE chooses an action maximizing the surrogate Lagrangian score,

$$a_t \in \arg\max_{k\in\{0,\dots,K\}} \left[\widetilde{r}_{k,t} - \lambda_t \widetilde{c}_{k,t}\right].$$

Therefore,

$$\widetilde{r}_{a_t,t} - \lambda_t \widetilde{c}_{a_t,t} \ge \sum_{k=0}^{K} \pi_k^\star(\phi_t)\left[\widetilde{r}_{k,t} - \lambda_t \widetilde{c}_{k,t}\right].$$

On the confidence event,

$$\widetilde{r}_{k,t} \ge \mu_{k,t}^r, \qquad \widetilde{c}_{k,t} \le \mu_{k,t}^c.$$

Since $\lambda_t \ge 0$,

$$\widetilde{r}_{a_t,t} - \lambda_t \widetilde{c}_{a_t,t} \ge R_t^\star - \lambda_t C_t^\star.$$

Rearranging,

$$\widetilde{r}_{a_t,t} \ge R_t^\star + \lambda_t(\widetilde{c}_{a_t,t} - C_t^\star).$$

Multiplying by I_t and summing over $t = 1, \dots, T$,

$$\sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} \ge \sum_{t=1}^{T} I_t R_t^\star + \sum_{t=1}^{T} I_t \lambda_t(\widetilde{c}_{a_t,t} - \rho) - \sum_{t=1}^{T} I_t \lambda_t (C_t^\star - \rho).$$

Using the benchmark concentration inequalities above gives

$$\sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} \ge \frac{\tau}{T}\mathrm{OPT} + \sum_{t=1}^{\tau} \lambda_t(\widetilde{c}_{a_t,t} - \rho) - (1+\Lambda)M_T(\delta).$$

□

Lemma 4 (Surrogate-to-realized cost bridge)**.** *On the joint confidence and concentration event,*

$$\sum_{t=1}^{\tau} \lambda_t(\widetilde{c}_{a_t,t} - \rho) \geq \sum_{t=1}^{\tau} \lambda_t(c_{a_t,t} - \rho) - \Lambda R_c(T) - \Lambda M_T(\delta).$$

Proof. We decompose

$$\sum_{t=1}^{\tau} \lambda_t(c_{a_t,t} - \widetilde{c}_{a_t,t}) = \sum_{t=1}^{\tau} \lambda_t(c_{a_t,t} - \mu^c_{a_t,t}) + \sum_{t=1}^{\tau} \lambda_t(\mu^c_{a_t,t} - \widetilde{c}_{a_t,t}).$$

The first term is a bounded weighted martingale-difference sum. Therefore, on the joint concentration event,

$$\sum_{t=1}^{\tau} \lambda_t(c_{a_t,t} - \mu^c_{a_t,t}) \leq \Lambda M_T(\delta).$$

For the second term, since $\lambda_t \leq \Lambda$ and

$$0 \leq \mu^c_{a_t,t} - \widetilde{c}_{a_t,t} \leq 2\alpha_c s_{a_t,t},$$

we obtain

$$\sum_{t=1}^{\tau} \lambda_t(\mu^c_{a_t,t} - \widetilde{c}_{a_t,t}) \leq \Lambda R_c(T).$$

Hence

$$\sum_{t=1}^{\tau} \lambda_t(c_{a_t,t} - \widetilde{c}_{a_t,t}) \leq \Lambda R_c(T) + \Lambda M_T(\delta),$$

which is equivalent to the claimed inequality. □

Theorem (Restatement of Theorem 1)**.** *Suppose the assumptions above hold. Let CARE use ridge parameter $\gamma > 0$, dual step size*

$$\eta = \Theta(\Lambda/\sqrt{T}),$$

and confidence parameters α_r, α_c chosen large enough to ensure the uniform self-normalized confidence event above. If

$$\Lambda \geq \frac{\mathrm{OPT}}{B},$$

then, with probability at least $1 - \delta$,

$$\mathrm{Reg}(T) = \widetilde{O}\left((1 + \Lambda) K d \sqrt{T}\right).$$

If, in addition, the dual cap satisfies the two-sided condition

$$\frac{\mathrm{OPT}}{B} \leq \Lambda \leq C_\Lambda \left(1 + \frac{\mathrm{OPT}}{B}\right)$$

for a universal constant C_Λ, then

$$\mathrm{Reg}(T) = \widetilde{O}\left(\left(1 + \frac{\mathrm{OPT}}{B}\right) K d \sqrt{T}\right).$$

Proof. Combining Lemma 3 and Lemma 4,

$$\sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} \geq \frac{\tau}{T}\mathrm{OPT} + \sum_{t=1}^{\tau} \lambda_t(c_{a_t,t} - \rho) - \Lambda R_c(T) - (1 + 2\Lambda) M_T(\delta).$$

Applying Lemma 2,

$$\sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} \geq \frac{\tau}{T}\mathrm{OPT} + \Lambda(B - \tau\rho) - \left(\frac{\Lambda^2}{2\eta} + \frac{\eta T}{2}\right) - \Lambda R_c(T) - (1+2\Lambda)M_T(\delta).$$

Since $\rho = B/T$,

$$\frac{\tau}{T}\mathrm{OPT} + \Lambda(B - \tau\rho) = \frac{\tau}{T}\mathrm{OPT} + \left(1 - \frac{\tau}{T}\right)\Lambda B.$$

Under

$$\Lambda \geq \frac{\mathrm{OPT}}{B},$$

we have

$$\Lambda B \geq \mathrm{OPT},$$

and therefore

$$\frac{\tau}{T}\mathrm{OPT} + \left(1 - \frac{\tau}{T}\right)\Lambda B \geq \mathrm{OPT}.$$

It follows that

$$\sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} \geq \mathrm{OPT} - \left(\frac{\Lambda^2}{2\eta} + \frac{\eta T}{2}\right) - \Lambda R_c(T) - (1+2\Lambda)M_T(\delta).$$

We now relate the optimistic reward surrogate to the realized reward. On the confidence event,

$$0 \leq \widetilde{r}_{a_t,t} - \mu^r_{a_t,t} \leq 2\alpha_r s_{a_t,t},$$

and hence

$$\sum_{t=1}^{\tau} (\widetilde{r}_{a_t,t} - \mu^r_{a_t,t}) \leq R_r(T).$$

The selected-reward martingale bound gives

$$\sum_{t=1}^{\tau} r_{a_t,t} \geq \sum_{t=1}^{\tau} \mu^r_{a_t,t} - M_T(\delta).$$

Combining the two inequalities,

$$\sum_{t=1}^{\tau} r_{a_t,t} \geq \sum_{t=1}^{\tau} \widetilde{r}_{a_t,t} - R_r(T) - M_T(\delta).$$

After τ, the theoretical process chooses the no-op action, whose reward is zero. Thus

$$\sum_{t=1}^{T} r_{a_t,t} = \sum_{t=1}^{\tau} r_{a_t,t}.$$

Therefore,

$$\mathrm{Reg}(T) = \mathrm{OPT} - \sum_{t=1}^{T} r_{a_t,t} \leq R_r(T) + \Lambda R_c(T) + \frac{\Lambda^2}{2\eta} + \frac{\eta T}{2} + O\left((1+\Lambda)M_T(\delta)\right).$$

Choosing

$$\eta = \Theta(\Lambda/\sqrt{T})$$

gives

$$\frac{\Lambda^2}{2\eta} + \frac{\eta T}{2} = O(\Lambda\sqrt{T}).$$

Furthermore,

$$R_r(T), R_c(T) = \widetilde{O}(Kd\sqrt{T}),$$

and

$$M_T(\delta) = \widetilde{O}(\sqrt{T}).$$

Hence

$$\mathrm{Reg}(T) = \widetilde{O}\left(Kd\sqrt{T} + \Lambda Kd\sqrt{T} + \Lambda\sqrt{T}\right),$$

which simplifies to

$$\mathrm{Reg}(T) = \widetilde{O}\left((1+\Lambda)Kd\sqrt{T}\right).$$

Finally, under the additional upper bound

$$\Lambda \le C_\Lambda\left(1 + \frac{\mathrm{OPT}}{B}\right),$$

we have

$$1 + \Lambda = O\left(1 + \frac{\mathrm{OPT}}{B}\right).$$

Therefore,

$$\mathrm{Reg}(T) = \widetilde{O}\left(\left(1 + \frac{\mathrm{OPT}}{B}\right)Kd\sqrt{T}\right).$$

□

C CARE ALGORITHM

Algorithm 1 summarizes the complete CARE procedure, including online reward and cost estimation, budget-aware expert selection, and the dual-variable update used to enforce the global computational budget.

Algorithm 1: CARE: Cost-Aware Online Routing for VLMs

Require: model pool $\mathcal{M} = \{m_k\}_{k=1}^K$, ridge parameter γ, exploration coefficients α_r, α_c, budget B, dual step size η, dual cap Λ, horizon T

$\lambda_1 \leftarrow 0$, $\rho \leftarrow B/T$, $B_{\mathrm{rem}} \leftarrow B$
for $k = 1, \dots, K$ **do**
  $A_k \leftarrow \gamma I_d$, $b_k^r \leftarrow 0$, $b_k^c \leftarrow 0$
end for
for $t = 1, \dots, T$ **do**
  observe x_t and compute $\phi_t = \mathcal{I}_{\mathrm{base}}(x_t)$
  for $k = 1, \dots, K$ **do**
    $\hat{\theta}_k^r \leftarrow A_k^{-1} b_k^r, \quad \hat{\theta}_k^c \leftarrow A_k^{-1} b_k^c$
    $\mathrm{UCB}_{k,t}^r \leftarrow \hat{\theta}_k^{r\top}\phi_t + \alpha_r\sqrt{\phi_t^\top A_k^{-1}\phi_t}$
    $\mathrm{LCB}_{k,t}^c \leftarrow \hat{\theta}_k^{c\top}\phi_t - \alpha_c\sqrt{\phi_t^\top A_k^{-1}\phi_t}$
    $\mathrm{Score}_{k,t} \leftarrow \mathrm{UCB}_{k,t}^r - \lambda_t \mathrm{LCB}_{k,t}^c$
  end for
  $a_t \leftarrow \arg\max_{k\in[K]} \mathrm{Score}_{k,t}$
  if $c_{a_t,t} > B_{\mathrm{rem}}$ **then**
    fallback to the base model
  else
    evaluate model m_{a_t} on x_t and observe $(r_{a_t,t}, c_{a_t,t})$
    $B_{\mathrm{rem}} \leftarrow B_{\mathrm{rem}} - c_{a_t,t}$
    $A_{a_t} \leftarrow A_{a_t} + \phi_t\phi_t^\top$
    $b_{a_t}^r \leftarrow b_{a_t}^r + r_{a_t,t}\phi_t$
    $b_{a_t}^c \leftarrow b_{a_t}^c + c_{a_t,t}\phi_t$
    $\lambda_{t+1} \leftarrow \Pi_{[0,\Lambda]}(\lambda_t + \eta(c_{a_t,t} - \rho))$
  end if
end for

D ADDITIONAL EXPERIMENTAL DETAILS

D.1 MODEL ZOO

Table 5 summarizes the model zoo used in our experiments. The pool includes both generalist and specialist VLMs spanning different architectures, pretraining corpora, and computational costs.

Model	Parameters	GFLOPs	Domain Specialty	Pretraining Data
CLIP ViT-B/32	151M	14.8	General image understanding	OpenAI
RemoteCLIP ViT-B/32	151M	14.8	Remote sensing and satellite imagery	Remote sensing corpora
BiomedCLIP ViT-B/16	150M	41.0	Biomedical and histopathology imagery	PMC-15M
ConvNeXt-Large-D	352M	107.5	Texture and local visual patterns	LAION2B-S26B-B102K-AugReg
CLIP ViT-L/14	428M	175.0	General image understanding	OpenAI
CLIP ViT-H/14@378	987M	1050.0	High-capacity general image understanding	DFN5B

Table 5: Model zoo used in our routing experiments. The pool includes both generalist and specialist experts spanning diverse computational costs, pretraining corpora, and application domains.

D.2 Prompt templates

For zero-shot classification, we use one text prompt per class and precompute text embeddings for each VLM.

CIFAR-10: `a photo of a {class}`
RESISC45: `a satellite image of {class}`
PCam:

- `a histopathology image of normal lymph node tissue`
- `a histopathology image of lymph node tissue containing metastatic tumor`

GTSRB: `a photo of a {class} traffic sign`
Food101: `a photo of {class}`
DTD: `a photo of a {class} texture`
Flowers102: `a photo of a {class} flower`
CRC: `a histopathology image of {class}`

D.3 Cost modeling

The general CARE formulation allows expert cost to be modeled online as a context-dependent quantity,

$$\mathbb{E}[c_{k,t} \mid \phi_t] = \theta_k^{c\top} \phi_t.$$

In our experiments, however, computational cost is defined from model FLOPs and is therefore deterministic and known before routing. We consequently use the known expert cost c_k directly rather than learning a separate contextual cost model.

Inference costs are derived from model FLOPs and normalized across experts using logarithmic scaling. The resulting quantity is used as a normalized routing-cost score rather than as a literal fraction of physical FLOPs. Since the base model is always evaluated to obtain the routing context, it is assigned zero *additional* cost; reported costs therefore measure only the extra expert computation beyond the base model.

D.4 Reward construction

CARE uses a bounded confidence-based proxy reward derived from the selected expert's predictive distribution. Let $p_{k,t}$ denote the softmax distribution obtained from the standardized logits of expert k for sample t. Our default reward is normalized negative entropy,

$$r_{k,t} = 1 - \frac{H(p_{k,t})}{\log C_t}, \qquad H(p_{k,t}) = -\sum_{j=1}^{C_t} p_{k,t,j} \log p_{k,t,j},$$

where C_t is the number of candidate classes for the current task. Since $0 \le H(p_{k,t}) \le \log C_t$, the resulting reward satisfies

$$r_{k,t} \in [0, 1].$$

This normalization provides a common bounded range across datasets with different numbers of classes and directly matches the bounded-reward assumption used in the theoretical analysis. Higher reward corresponds to lower predictive entropy and therefore greater model confidence. Importantly, this quantity is a label-free confidence proxy rather than a correctness signal. The theoretical guarantee therefore applies directly to this bounded default reward. Alternative confidence functions considered in the ablation study are evaluated empirically and need not satisfy the same bounded-reward assumptions unless appropriately normalized.

Logit standardization details. For numerical stability, standardized logits are computed as

$$\tilde{\ell} = \frac{\ell - \mu(\ell)\mathbf{1}}{\sigma(\ell) + \epsilon}, \qquad \epsilon = 10^{-6}.$$

Accordingly, the affine-invariance property in Proposition 1 is exact when $\epsilon = 0$ and is approximately satisfied when $\epsilon \ll \sigma(\ell)$. For binary classification, the nonzero regularizer also prevents exact confidence collapse by retaining a dependence on the original logit margin.

Mini-batched updates. The theoretical analysis in Appendix B considers the idealized sequential setting in which routing statistics and the dual variable are updated after every round. The experimental implementation instead performs updates in mini-batches for computational efficiency. Decisions within a batch use the routing parameters available at the beginning of that batch, and the accumulated observations are used for the subsequent update. Sensitivity to batch size is reported in Appendix G.

D.5 Baselines

Confidence oracle. For each input x_t, the Confidence Oracle evaluates all experts $\{m_k\}_{k=1}^K$ and selects the prediction from the model with the highest confidence score:

$$k^* = \arg\max_k u_k(x_t), \quad \hat{y}_t = \arg\max_j \ell_{k^*,j}(x_t),$$

where $u_k(x_t)$ denotes the confidence score computed from the model logits. This baseline provides a full-information confidence-based reference, incurring the cost of evaluating all experts for every sample.

Weighted ensemble. The Weighted Ensemble combines predictions from all experts by computing a weighted average of their logits:

$$\bar{\ell}(x_t) = \sum_{k=1}^{K} w_k(x_t)\, \ell_k(x_t), \quad w_k(x_t) = \frac{\exp(u_k(x_t))}{\sum_{i=1}^{K} \exp(u_i(x_t))}.$$

The final prediction is $\hat{y}_t = \arg\max_j \bar{\ell}_j(x_t)$. This baseline also incurs the full cost of evaluating all experts.

Fixed cascade. The Fixed Cascade processes models sequentially in increasing cost order. Let $\{m_{(1)}, \ldots, m_{(K)}\}$ denote models sorted by cost. For each input x_t, we evaluate models in order and stop at the first model whose confidence exceeds a fixed threshold τ:

$$\text{select } m_{(k)} \text{ such that } u_{(k)}(x_t) \geq \tau,$$

otherwise proceeding to the next model. The threshold τ is determined from the first batch as the 60th percentile of base model confidence scores and is kept fixed thereafter.

Neural-LinUCB. Following Xu et al. (2022), we implement a Neural-LinUCB baseline with a two-layer MLP of hidden dimension 64. UCB exploration is applied only to the learned last-layer representation to keep online uncertainty estimation tractable. The method uses the same label-free confidence reward as CARE and requires no separate reference or labeled calibration set. We do not impose an explicit global cost constraint on this baseline.

PORT-inspired router. Inspired by PORT Wu & Silwal (2025), we construct a nonparametric nearest-neighbor router using 256 unlabeled reference samples. All experts are evaluated offline on the reference set, and we store the base-VLM representation and expert confidence scores for each sample. At test time, an input is routed to the expert with the highest average confidence among its nearest reference neighbors. This baseline therefore requires an unlabeled reference set and offline full-expert profiling.

Supervised calibrated cascade. We reserve 10% of each dataset as labeled calibration data and fit one temperature per VLM by minimizing negative log-likelihood. Models are ordered by increasing cost, and stage-wise exit thresholds are optimized over model-specific confidence quantiles at 5% increments for three coordinate-wise sweeps, inspired by Zellinger & Thomson (2025). The temperatures and thresholds are then frozen before evaluation. Unlike CARE, this baseline requires labeled pre-deployment data and offline threshold optimization.

D.6 COMPUTE

All experiments are conducted on a server equipped with an NVIDIA GeForce RTX 4090 GPU. We use PyTorch 2.9.1 with CUDA 12.8 for model inference. Unless otherwise stated, experiments are run with a batch size of 128. A full routing sweep across all datasets takes approximately 2 hours, highlighting the practical efficiency of the proposed method.

E DETAILED CONFIDENCE METRIC COMPARISON

Table 6 reports a detailed comparison of several confidence (reward) metrics computed from both raw logits and standardized (z-score normalized) logits across all datasets. We evaluate maximum logit, maximum softmax probability, entropy, and energy for both oracle selection and CARE routing. The results show that confidence metrics based on raw logits are highly unreliable due to inter-model scale differences, often leading to poor performance. In contrast, using standardized logits consistently improves accuracy for both oracle and routing, yielding more stable and comparable confidence estimates across models.

	CIFAR-10		RESISC45		PCam		GTSRB		Food101		DTD		Flowers102		CRC	
Metric	Raw	Norm	Raw	Norm	Raw	Norm	Raw	Norm	Raw	Norm	Raw	Norm	Raw	Norm	Raw	Norm
Oracle Accuracy (%)																
Max Logit	33.4	95.8	17.0	77.2	75.6	75.9	13.9	50.0	3.3	95.4	11.2	63.5	1.1	89.6	42.4	40.9
Max Softmax	95.3	95.7	74.1	77.5	75.9	76.0	56.0	49.5	80.1	95.4	64.6	63.1	88.1	89.4	43.0	40.8
Entropy	79.0	95.9	55.7	75.2	76.0	76.0	36.8	51.4	2.1	95.3	43.1	64.1	78.4	89.9	42.7	40.9
Energy	28.0	96.1	16.6	74.5	75.5	76.0	7.8	51.7	5.5	95.2	10.9	64.6	0.9	90.4	33.7	41.1
CARE Routing Accuracy (%)																
Max Logit	33.5	95.4	18.4	68.3	75.0	66.5	15.7	45.8	2.9	91.5	16.9	53.8	4.6	81.0	39.2	38.2
Max Softmax	83.1	95.4	66.6	67.6	65.2	66.7	27.6	45.8	42.0	91.4	44.9	53.7	77.1	80.2	39.8	38.9
Entropy	62.0	95.2	47.3	68.7	66.2	66.5	4.0	46.4	2.8	91.2	18.9	54.1	75.1	81.1	39.2	35.5
Energy	38.5	95.4	18.6	68.7	75.0	66.3	7.9	45.8	37.2	91.3	15.2	54.0	13.4	80.5	26.3	34.9

Table 6: Per-dataset comparison of confidence metrics for oracle selection and CARE routing under raw and normalized logits, with CARE evaluated at a target average budget of $\rho = 0.6$. Normalized confidence generally improves performance by correcting for inter-model logit scale differences.

F STABILITY UNDER RANDOMIZED TEST STREAMS

Since CARE operates in an online manner and depends on the order of incoming samples, we evaluate robustness to input ordering by running each experiment three times with different random shuffles of the test stream. Table 7 reports the mean and standard deviation of routing accuracy for various ρ. Across all datasets, the standard deviation remains low (typically below 1%), indicating that CARE is robust to variations in the input stream. Notably, even on more challenging datasets such as PCam and CRC, the variance remains moderate, suggesting that CARE consistently maintains performance despite stochastic differences in sample ordering.

	CIFAR-10	RESISC45	PCam	GTSRB	Food101	DTD	Flowers102	CRC
CARE-20	91.56 ± 0.06	61.33 ± 0.44	61.75 ± 0.93	31.24 ± 0.25	84.00 ± 0.02	41.32 ± 1.03	68.55 ± 0.43	26.82 ± 1.57
CARE-60	95.45 ± 0.08	70.55 ± 0.52	65.03 ± 1.08	46.47 ± 0.21	91.27 ± 0.02	53.19 ± 0.83	82.17 ± 0.37	41.06 ± 1.33
CARE-100	95.72 ± 0.05	70.78 ± 0.68	65.41 ± 1.23	51.65 ± 0.35	95.46 ± 0.01	57.09 ± 0.70	88.31 ± 0.24	42.11 ± 1.43

Table 7: Routing accuracy (%) reported as mean $\pm$ standard deviation over three runs with different random shuffles of the test stream.

G HYPERPARAMETER SENSITIVITY

Tables 8, 9, and 10 examine the sensitivity of CARE to the exploration coefficient α_r, dual update step size η, and batch size under $\rho = 0.6$. CARE remains relatively stable across a broad range of α_r, with moderate values generally performing best. Performance and budget tracking are also stable for small-to-moderate values of η, while larger updates can slightly reduce accuracy and increase budget error. Finally, smaller and moderate batch sizes generally

benefit online adaptation through more frequent updates, although the effect is not strictly monotonic. Overall, CARE does not require highly sensitive hyperparameter tuning.

α_r	CIFAR-10	RESISC45	PCam	GTSRB	Food101	DTD	Flowers102	CRC	Avg.
0.10	96.28	68.57	62.94	46.63	90.36	52.87	76.53	37.13	66.41
0.20	96.44	69.40	62.61	46.25	91.21	53.56	77.30	35.26	66.50
0.50	95.74	68.98	65.84	46.34	91.10	53.88	83.77	35.88	67.69
0.75	95.41	68.94	66.83	46.32	91.49	53.94	82.35	35.43	67.59
1.00	95.24	68.65	66.55	46.39	91.25	54.10	81.14	35.54	67.36
1.50	95.18	68.19	66.02	46.25	90.94	54.26	79.90	37.41	67.27
2.00	95.23	68.11	66.98	45.99	90.67	53.30	78.97	36.03	66.91

Table 8: Sensitivity to the exploration coefficient α_r at $\rho = 0.6$.

η	CIFAR-10	RESISC45	PCam	GTSRB	Food101	DTD	Flowers102	CRC	Avg. Acc.	Avg. Budget Error
0.01	94.56	67.25	66.06	46.71	89.07	52.93	82.26	37.09	66.99	0.0298
0.05	94.77	67.76	65.68	46.56	90.44	52.77	81.96	36.73	67.08	0.0296
0.10	94.78	67.95	66.23	46.26	91.01	52.50	80.53	37.17	67.05	0.0295
0.50	95.24	68.65	66.55	46.39	91.25	54.10	81.14	35.54	67.36	0.0232
1.00	95.33	69.03	66.25	46.22	91.03	52.71	79.96	34.65	66.90	0.0371
2.00	95.26	68.83	65.16	45.87	90.11	50.85	79.96	35.32	66.42	0.0577

Table 9: Sensitivity to the dual update step size η at $\rho = 0.6$. The final column reports the average budget error.

Batch Size	CIFAR-10	RESISC45	PCam	GTSRB	Food101	DTD	Flowers102	CRC	Avg.
8	97.25	71.06	64.46	47.28	91.26	59.68	79.54	35.11	68.21
16	96.82	72.81	64.11	46.84	91.11	58.67	80.78	33.40	68.07
32	96.84	73.35	66.42	46.61	91.26	57.39	79.54	37.81	68.65
64	96.18	70.57	66.03	46.06	91.49	53.78	80.66	35.24	67.50
128	95.24	68.65	66.55	46.39	91.25	54.10	81.14	35.54	67.36

Table 10: Sensitivity to batch size at $\rho = 0.6$.

H Mixed-Domain Online Routing

The preceding experiments evaluate each dataset as a separate online stream. To evaluate CARE on a heterogeneous stream, we randomly interleave samples from all eight datasets. CARE processes the resulting stream without resetting its reward and cost estimators or dual variable. Dataset or domain identity is not provided as an input to the routing policy; the task-specific class vocabulary is supplied to compute the corresponding zero-shot logits. Table 11 shows that, under $\rho = 0.6$, CARE achieves an average accuracy of 67.00% on the mixed stream, compared with 67.36% when the datasets are processed separately. These results support CARE's ability to route across interleaved domains without explicit domain identifiers as routing inputs.

I Routing Overhead and Inference Latency.

To quantify CARE's runtime overhead, we measure per-sample inference latency on CIFAR-10 with batch size 128. We separately report VLM inference latency and the additional CARE routing overhead, which includes UCB evaluation and online linear updates. As shown in Table 12, the router adds only 0.0814 ms per sample, compared with 0.3541 ms for the base model and substantially larger latencies for the high-capacity experts. Thus, CARE introduces limited additional runtime while enabling adaptive expert selection.

Setting	CIFAR-10	RESISC45	PCam	GTSRB	Food101	DTD	Flowers102	CRC	Avg.
Separate	95.24	68.65	66.55	46.39	91.25	54.10	81.14	35.54	67.36
Mixed-domain	95.35	67.23	64.13	45.79	91.30	52.19	82.34	38.13	67.00

Table 11: Comparison of CARE on separate and mixed-domain streams.

Component	Latency (ms/sample)
Base ViT-B/32	0.3541
CARE Router	0.0814
RemoteCLIP	0.3482
BiomedCLIP	1.1174
ConvNeXt-Large-D	3.1839
CLIP ViT-L/14	5.0103
CLIP ViT-H/14-378	32.6393

Table 12: Inference latency and CARE routing overhead.

J ACCURACY–COST DYNAMICS OVER BATCHES FOR DIFFERENT BUDGET LEVELS

This section presents cumulative accuracy and average cost over routing batches for different budget levels. While early batches may show variability due to exploration, both metrics stabilize as more samples are processed. The average cost consistently tracks the target budget ρ, demonstrating stable and budget-aware online routing.

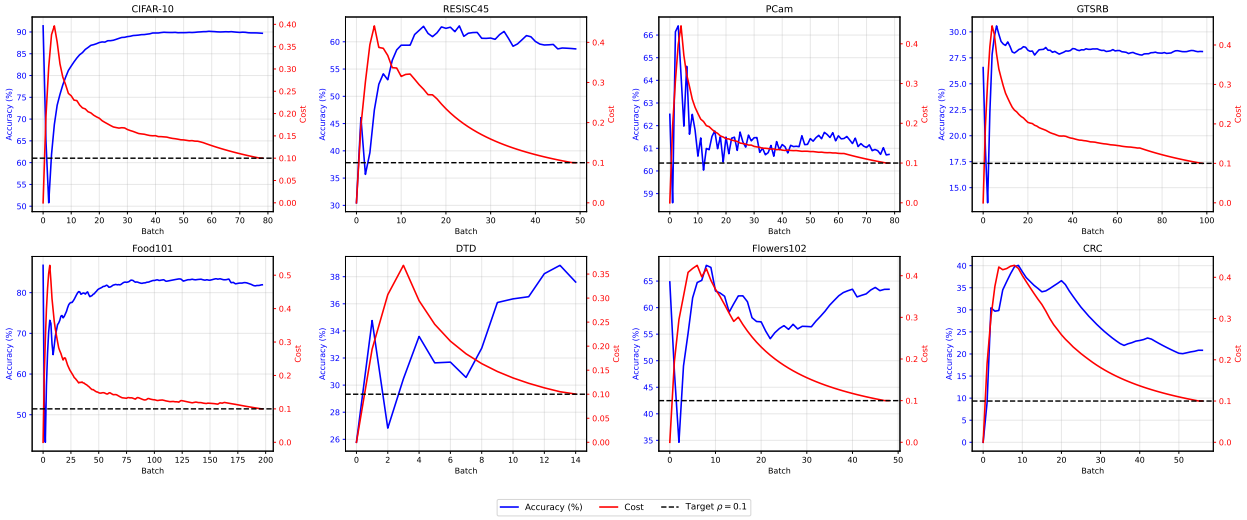

Figure 4: Cumulative accuracy and average cost over routing rounds for $\rho = 0.1$.

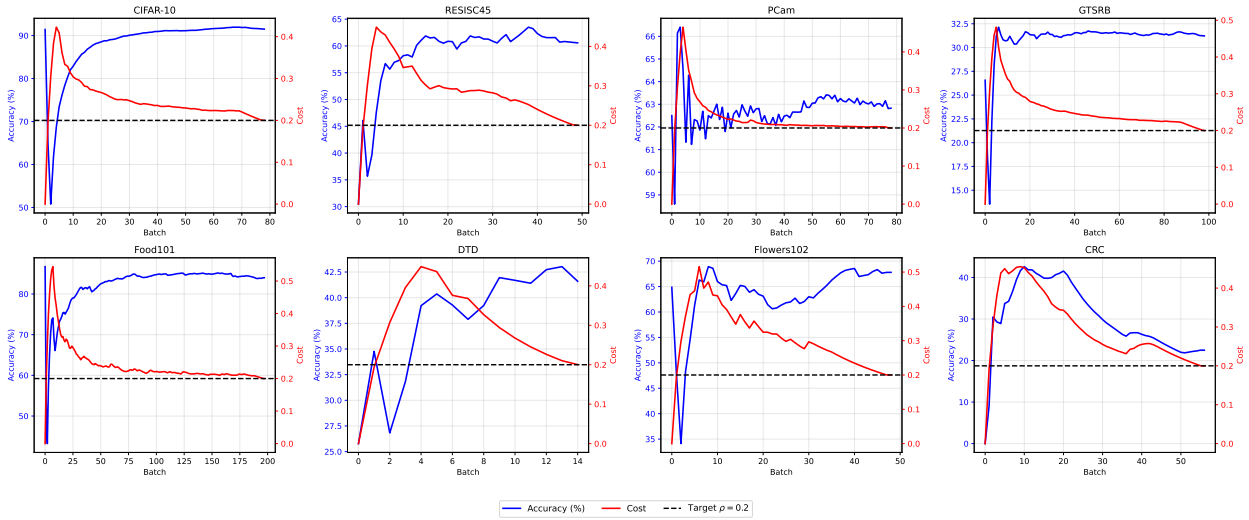

Figure 5: Cumulative accuracy and average cost over routing rounds for $\rho = 0.2$.

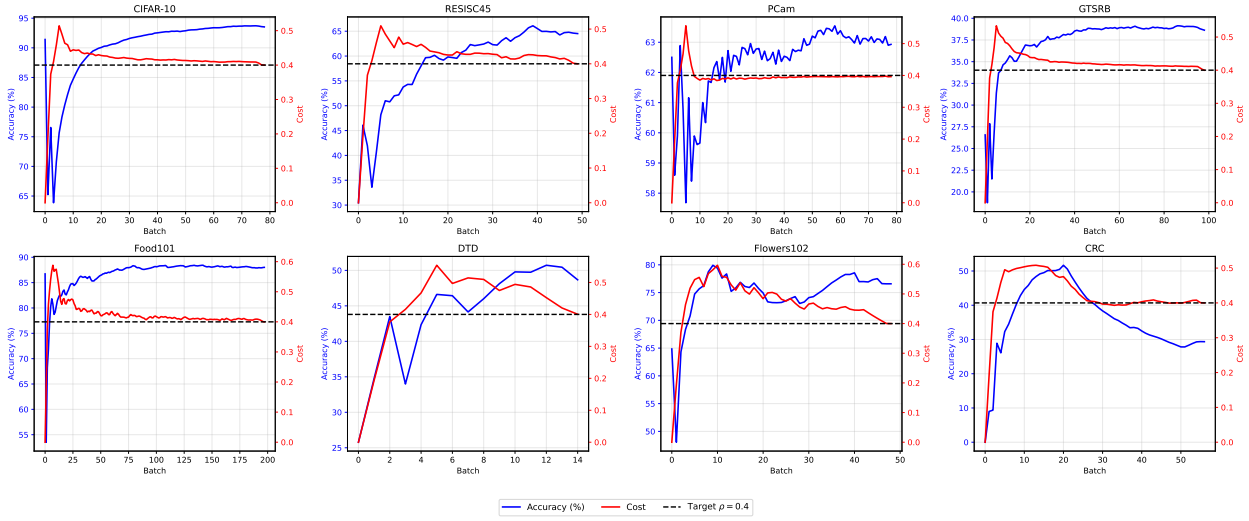

Figure 6: Cumulative accuracy and average cost over routing rounds for $\rho = 0.4$.

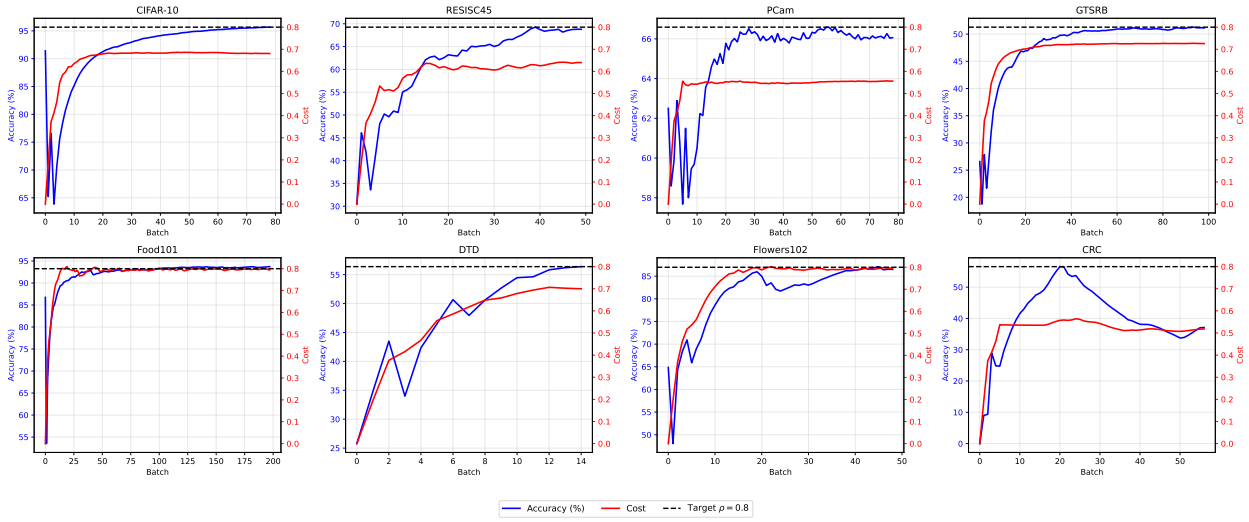

Figure 7: Cumulative accuracy and average cost over routing rounds for $\rho = 0.8$.

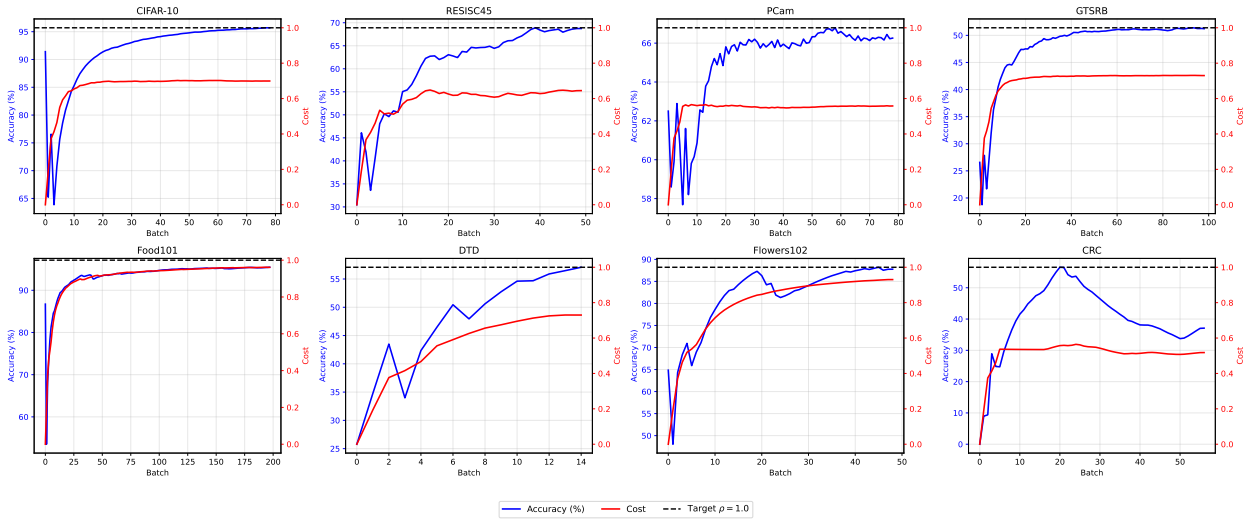

Figure 8: Cumulative accuracy and average cost over routing rounds for $\rho = 1.0$.