CONFORMAL ROUTING OF PARTIALLY TRANSFERABLE EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transferability need not be a property of an entire source model; useful computations can coexist with source-specific shortcuts, and their value can vary across target inputs. We introduce ROUTECP (*Conformal Routing of Partially Transferable Experts*), which aligns reusable source experts to a target representation and learns a sparse input-dependent route over them. Training directly targets conformal efficiency under a compute budget, while a finite-partition safety certificate suppresses expert contributions that cannot be certified as nonharmful. An untouched target calibration split forms a calibration firewall, preserving finite-sample marginal target coverage under calibration-test exchangeability despite adaptive source training, routing, selection, and certification. We derive an oracle inequality for sparse local routing, prove a separation from whole-source transfer under partial sharing, and establish simultaneous cell-average safety over a finite target partition. Simulations and real-data experiments show smaller prediction sets, improved low-resource transfer, reduced negative transfer, and favorable compute-efficiency tradeoffs at nominal coverage across heterogeneous target domains.

1 INTRODUCTION

Transfer learning typically treats a source model as an indivisible unit: pool it, fine-tune it, weight it, or discard it. This granularity can be wrong when different components have different transfer behavior. A source encoder may contain broadly useful low-level representations, semantic modules useful only for selected target inputs, and shortcuts that become harmful under shift. Full sharing imports useful and harmful computations together, target-only learning discards reusable structure, and whole-source selection cannot represent partial, input-specific usefulness. The relevant question is thus *which source computations should be reused for a particular target input?* Figure 1 summarizes the key insight.

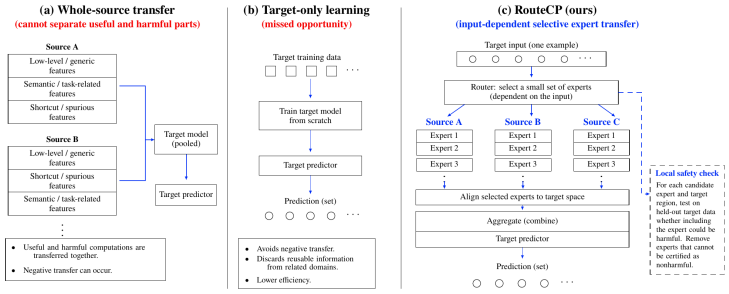


Figure 1: **RouteCP in one view.** Source models expose reusable experts. A target-conditioned sparse router selects a small subset, transports them into the target space, and applies a local nonharmfulness certificate. Final conformal calibration is isolated from all adaptive transfer decisions.

Conformal prediction gives finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Shafer and Vovk, 2008; Lei et al., 2018; Angelopoulos and Bates, 2023). Recent work develops adaptive prediction sets, conformalized training, and methods for distribution shift (Tibshirani et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021; Bates et al., 2021; Koh et al., 2021; Einbinder et al., 2022; Kasa et al., 2025; Liu et al., 2024; Plassier et al., 2024; Xu et al., 2025; Zhang et al., 2026). These approaches largely modify calibration weights, training observations, or source/domain contributions, but do not jointly model transferability across both reusable representation components and individual

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.

target inputs. Modern architectures, meanwhile, expose natural transfer units such as residual blocks, adapters, low-rank updates, attention heads, and mixture-of-experts components (Shazeer et al., 2017; Houlsby et al., 2019; Pfeiffer et al., 2021; Riquelme et al., 2021; Fedus et al., 2022; Hu et al., 2022). Domain-adaptation and negative-transfer studies further show that source relatedness can vary substantially under shift (Long et al., 2018; Peng et al., 2019; Wang et al., 2019; Gulrajani and Lopez-Paz, 2021). The missing object is therefore a rule that decides *which reusable source computations to activate for each target input* while retaining a separate statistical mechanism for predictive reliability.

We introduce ROUTECP, which aligns source experts to a common target space, learns an input-dependent sparse route, and trains the resulting predictor for predictive and conformal efficiency under explicit computation constraints. A held-out target sample then certifies expert-region contributions before an untouched calibration sample determines the final conformal threshold. The certificate and conformal calibration serve different purposes; certification controls whether transferred computations are retained, whereas calibration controls coverage of the resulting prediction set. Thus negative-transfer protection does not replace or modify the final exchangeable-rank guarantee.

Our contributions are threefold. First, we formulate transferability at the level of input-dependent reusable computations instead of entire source models, yielding a compute-aware sparse router that explicitly includes the target-only action. Second, we separate route learning, model selection, independent safety certification, and final conformal calibration. For this procedure, an oracle inequality quantifies sparse-route, source/transport, selection, and optimization errors, while a separation theorem constructs partial-sharing problems on which every whole-source rule has nonvanishing regret but local routing is consistent. Third, a finite-partition certificate provides simultaneous cell-average control of harmful retained routes without being required for marginal conformal validity. Experiments examine route recovery, negative-transfer protection, low-resource scaling, and computation. Full model assumptions, technical proofs, optimization details, and extended results appear in the appendix.

2 PARTIAL TRANSFER AND LOCAL EXPERT ROUTING

We observe S labeled sources $\mathcal{D}_s = \{(X_{si}, Y_{si})\}_{i=1}^{n_s} \sim P_s^{n_s}$ sharing labels $[C] = \{1, \dots, C\}$, and target distribution P_0. Before fitting, target labels are split into disjoint $\mathcal{D}_0^{\mathrm{tr}}, \mathcal{D}_0^{\mathrm{sel}}, \mathcal{D}_0^{\mathrm{cal}}$, with $\mathcal{D}_0^{\mathrm{sel}} = \mathcal{D}_0^{\mathrm{val}} \dot{\cup} \mathcal{D}_0^{\mathrm{cert}}$ fixed in advance. Their sizes are $n_{\mathrm{tr}}, n_{\mathrm{val}}, n_{\mathrm{cert}}$, and n_{cal}. Only $\mathcal{D}_0^{\mathrm{cal}}$ is used for the final conformal rank, preserving strict separation from adaptation. Conditional on all preceding adaptive choices, coverage requires only exchangeability of $\mathcal{D}_0^{\mathrm{cal}}$ and a future $(X_*, Y_*) \sim P_0$; no source-target relation is required for validity, despite arbitrary source-domain distributional mismatch. We present multiclass classification; Appendix D gives the corresponding regression construction.

Source s exposes K_s prespecified expert maps $h_{sk} : \mathcal{X} \to \mathbb{R}^{d_{sk}}$, while $h_0 : \mathcal{X} \to \mathbb{R}^d$ denotes the target backbone. We write $e = (s, k)$, $\mathcal{E} = \{(s, k) : s \in [S],\ k \in [K_s]\}$, $M = |\mathcal{E}|$, and $d_e = d_{sk}$ for $e = (s, k)$. Because source representations need not be commensurate, expert e is mapped into the target space as $\bar{h}_e(x) = \mathrm{LN}\{T_e(h_e(x))\}$, where $T_e : \mathbb{R}^{d_e} \to \mathbb{R}^d$ is the affine transport $T_e(v) = W_e v + b_e$, with $W_e \in \mathbb{R}^{d \times d_e}$ and $b_e \in \mathbb{R}^d$. Here LN denotes layer normalization. We use affine transports throughout the theoretical analysis and empirical study. Their parameters are estimated using only source and noncalibration target data. The architectural partition is fixed before target model selection; in our experiments an expert is a residual stage, adapter block, grouped attention module, or LoRA update. This prevents target validation data from being used to manufacture a favorable decomposition. Layer normalization is applied after transport so that routing weights cannot be driven merely by arbitrary differences in expert scale. The role of T_e is limited to affine coordinate alignment and excludes unrestricted target-side representation learning; transferred features must therefore remain functions of the corresponding source expert h_e. This restriction makes the transfer unit auditable.

The distinction between *proposing* and *evaluating* experts is computationally important. A router that first computes every $h_e(x)$ could make a more informed decision, but it would incur dense-source inference cost before sparsification (Zhou et al., 2022). Our default router therefore conditions on $h_0(x)$ and learned expert identities u_e, chooses a support, and only then evaluates the selected source modules. The expert embeddings are learned jointly with the router and can encode source identity, architectural location, and persistent target relevance, whereas the dependence on $h_0(x)$ allows the same expert to be activated for one target input and suppressed for another. A richer router could pre-evaluate source outputs, but doing so would sacrifice the sparse-evaluation advantage; the main method therefore conditions only on the target representation and expert identities before support selection.

Each expert has embedding $u_e \in \mathbb{R}^{d_u}$. From the target representation, the router forms $r_e(x) = a^\top \sigma\{A_0 h_0(x) + A_u u_e + c\}$ and selects the q largest logits $I_q(x)$, with ties resolved by a fixed deterministic rule, where σ is a coordinatewise activation function and $\phi = \{a, A_0, A_u, c, (u_e)_{e\in\mathcal{E}}\}$ collects the router parameters. With temperature $\tau > 0$, $\pi_e(x) = \exp\{r_e(x)/\tau\}\mathbf{1}\{e \in I_q(x)\}/\sum_{j\in I_q(x)} \exp\{r_j(x)/\tau\}$, so $\pi_e(x) \geq 0$, $\sum_e \pi_e(x) = 1$, and $\|\pi(x)\|_0 \leq q$. This compute-aware router chooses support before evaluating source experts. The routed representation and probabilities are

$$z(x) = h_0(x) + \gamma \sum_e \pi_e(x)\bar{h}_e(x), \quad p(x) = \mathrm{softmax}\{g_\theta(z(x))\}, \quad \gamma \in [0, \gamma_{\max}], \tag{2.1}$$

where a prediction head $g_\theta : \mathbb{R}^d \to \mathbb{R}^C$ produces logits. Thus $\gamma = 0$ nests target-only prediction; certification also falls back to h_0 if every expert is removed. This nesting is more than a convenience; every candidate class contains a no-transfer action, so additional sources cannot force the deployed predictor to use transferred computation. The learned value of γ controls the overall transfer amplitude, whereas $\pi_e(x)$ determines how that amplitude is distributed among locally selected experts. Keeping these roles separate avoids conflating "how much to transfer" with "what to transfer" and gives the safety stage a well-defined target-only counterfactual.

Let $\bar{h}_\pi = \sum_e \pi_e \bar{h}_e$ and $d^2(x) = \sum_e \pi_e(x)\|\bar{h}_e(x) - \bar{h}_\pi(x)\|_2^2$. For least-ambiguous-set classification (LAC) we use $V_\eta(x, y) = \{1 - p_y(x)\}\{1 + \eta d(x)\}$, $\eta \geq 0$, where $p_y(x)$ is the probability assigned to candidate label y; APS/RAPS multiply their cumulative score by the same factor (Angelopoulos et al., 2024). At deployment, both $p(x)$ and $d(x)$ are recomputed using the certified route after deletion and renormalization; if no expert survives, they reduce to their target-only counterparts. Disagreement adaptation is optional and affects efficiency, not validity, because η is fixed before calibration. The quantity $d(x)$ is computed only from experts that survive certification at deployment. It therefore summarizes residual disagreement among the computations the procedure is actually willing to use, instead of disagreement in the entire source bank. The factor $1+\eta d(x)$ increases nonconformity when retained experts disagree; after calibration, its effect on set size is determined by the resulting score distribution. We do not interpret $d(x)$ as a calibrated epistemic uncertainty estimate; its role is an observable routing diagnostic whose contribution is selected on validation data and isolated by ablation.

For efficiency analysis only, let $f_0^* : \mathcal{X} \to \mathbb{R}^C$ be the centered target logit and $f_e^* \in \mathcal{G}_e$ the target-space contribution induced by source expert e. We assume

$$f_0^*(x) = f_{\mathrm{private}}^*(x) + \sum_{e\in\mathcal{E}} a_e^*(x) f_e^*(x), \quad \|a^*(x)\|_0 \leq q, \quad \|a^*(x)\|_\infty \leq A_{\max}, \tag{2.2}$$

where $a^*(x) = \{a_e^*(x) : e \in \mathcal{E}\}$ collects input-dependent scalar coefficients and $A_{\max} < \infty$ is a fixed uniform bound. These oracle coefficients describe population partial sharing and need not coincide with the learned normalized weights $\pi_e(x)$; any mismatch between the structural representation and the nonnegative routed class is absorbed by the approximation error $\mathcal{A}(\Pi_q)$ defined below. The model is restrictive in that each contribution belongs to its source-induced class, coefficients are uniformly bounded, and at most q experts are locally active. It includes global module sharing when $a_e^*(x)$ is constant in x, and whole-source sharing when all experts from an accepted source are activated jointly. The module-level decomposition need not be uniquely identifiable because our theoretical target is the induced logit and conformal risk, not a semantic interpretation of individual neural components. This avoids imposing unverifiable assumptions about what a particular attention head or adapter "means."

For a fixed candidate architecture and parameter bounds, let Π_q be the induced router class and $\mathcal{F}(\Pi_q)$ the corresponding routed-logit class, allowing the admissible target backbone, prediction head, transports, router parameters, and $\gamma \in [0, \gamma_{\max}]$ to vary. The approximation error $\mathcal{A}(\Pi_q) = \inf_{f\in\mathcal{F}(\Pi_q)} \{E_{P_0}\|f(X) - f_0^\star(X)\|_2^2\}^{1/2}$ separates misspecification of the chosen expert partition and transport family from finite-sample learning error. In particular, the theory does not assume that the oracle route is exactly representable unless $\mathcal{A}(\Pi_q) = 0$. Appendix A gives the formal function classes and entropy conditions. Appendix D gives the regression analogue.

3 Conformal Routing

We split $\mathcal{D}_0^{\mathrm{sel}} = \mathcal{D}_0^{\mathrm{val}} \dot{\cup} \mathcal{D}_0^{\mathrm{cert}}$ in advance. No quantity below uses $\mathcal{D}_0^{\mathrm{cal}}$ until the last step. Training on $\mathcal{D}_0^{\mathrm{tr}}$ minimizes

$$\hat{L} = \hat{L}_{\mathrm{ce}} + \lambda_{\mathrm{cp}}\hat{L}_{\mathrm{conf}} + \lambda_{\mathrm{al}}\hat{L}_{\mathrm{align}} + \lambda_{\mathrm{ent}}\hat{L}_{\mathrm{ent}} + \lambda_{\mathrm{bal}}\hat{L}_{\mathrm{bal}}, \tag{3.1}$$

where $\hat{L}_{\mathrm{ce}}$ is the ordinary target cross-entropy loss and $\hat{L}_{\mathrm{conf}}$ directly approximates conformal set size. Specifically, with training scores $v_i = V_\eta(X_i, Y_i)$, we use the unique smoothed quantile $\tilde{q}_{1-\alpha} = \arg\min_t \{n_{\mathrm{tr}}^{-1} \sum_i \rho_{1-\alpha,h_q}(v_i - t) + \xi_q t^2/2\}$, where ρ_{u,h_q} is the Huber-smoothed pinball loss at quantile level u, $h_q > 0$ is its smoothing bandwidth, and $\xi_q > 0$ is a fixed regularization constant ensuring a unique minimizer, and for LAC, $\hat{L}_{\mathrm{conf}} = n_{\mathrm{tr}}^{-1} \sum_i \sum_c s\{(\tilde{q}_{1-\alpha} - V_\eta(X_i, c))/h_{\mathrm{set}}\}$, where s is logistic and $h_{\mathrm{set}} > 0$ controls the smooth approximation to set membership. This is only a training surrogate; validation and calibration use ordinary empirical quantiles. The alignment term matches class-conditional source/target means together with their marginal covariance structure,

$$\hat{L}_{\mathrm{align}} = \sum_{e=(s,k)} \Big\{ \sum_{c=1}^{C} \|\hat{\mu}_{e,c}^{(s)} - \hat{\mu}_c^{(0)}\|_2^2 + \kappa \|\hat{\Sigma}_e^{(s)} - \hat{\Sigma}^{(0)}\|_F^2 \Big\}. \tag{3.2}$$

For expert $e = (s, k)$, $\hat{\mu}_{e,c}^{(s)}$ and $\hat{\Sigma}_e^{(s)}$ are the empirical mean and covariance of the transported source representation $\bar{h}_e(X)$ among source-s observations with label c, whereas $\hat{\mu}_c^{(0)}$ and $\hat{\Sigma}^{(0)}$ are the corresponding target-backbone moments computed from $\mathcal{D}_0^{\mathrm{tr}}$. The coefficient $\kappa \geq 0$ determines the relative contribution of second-moment alignment. The entropy and balancing terms are $\hat{L}_{\mathrm{ent}} = n_{\mathrm{tr}}^{-1} \sum_i H\{\pi(X_i)\}$, where $H(\pi) = -\sum_{e \in \mathcal{E}} \pi_e \log \pi_e$, and $\hat{L}_{\mathrm{bal}} = M \sum_{e \in \mathcal{E}} (\bar{\pi}_e - M^{-1})^2$, where $\bar{\pi}_e = n_{\mathrm{tr}}^{-1} \sum_i \pi_e(X_i)$. They promote concentrated but noncollapsed routing. Forward passes use hard top-q support; backward passes differentiate the masked softmax while treating the support fixed. We therefore retain an explicit optimization-error term without assuming global optimization. The five terms in (3.1) have distinct roles; cross-entropy fits the target decision rule, $\hat{L}_{\mathrm{conf}}$ favors smaller sets at the selected score quantile, alignment makes source experts comparable in the target space, entropy avoids diffuse routes, and balancing prevents a few experts from absorbing the full target population. Smoothing derivatives and the alternating optimizer are in Appendix B.

A finite grid Λ of route budgets, regularization, score parameters, and smoothing constants is fixed before $\mathcal{D}_0^{\mathrm{val}}$ is inspected. Each candidate is trained on $\mathcal{D}_0^{\mathrm{tr}}$. We partition $\mathcal{D}_0^{\mathrm{val}}$ into fixed folds; for each fold, the conformal threshold is computed from the remaining validation folds and evaluated on the held-out fold, with the fitted candidate itself kept fixed. The resulting foldwise conformal evaluation chooses the smallest mean set size satisfying validation coverage $\geq 1 - \alpha - \epsilon_{\mathrm{sel}}$ and a prespecified FLOP budget, where $\epsilon_{\mathrm{sel}} \geq 0$ is fixed in advance. This constraint is only a selection safeguard, not part of the final coverage theorem. It also separates the optimized surrogate from selection; candidates are ranked using ordinary conformal sets on held-out validation folds, not the smoothed quantile used for gradients. The selected specification is refit once on $\mathcal{D}_0^{\mathrm{tr}} \cup \mathcal{D}_0^{\mathrm{val}}$; no refit occurs after certification. This order avoids the pitfall of certifying one fitted router and deploying a different refitted router.

For safety, we construct a finite target partition $\mathcal{G} = \{G_j\}_{j=1}^J$ from the refitted model without $\mathcal{D}_0^{\mathrm{cert}}$ or $\mathcal{D}_0^{\mathrm{cal}}$. Cells use only deployable quantities (predicted class, target-backbone confidence, and disagreement bins), with boundaries fixed from the fit sample. Using a reference threshold $\bar{q}$ computed from the selected candidate and its validation scores, and therefore fixed before either $\mathcal{D}_0^{\mathrm{cert}}$ or $\mathcal{D}_0^{\mathrm{cal}}$ is accessed, define

$$\ell_{\mathrm{safe}}(x, y; \pi) = \frac{1}{C} \sum_{c=1}^{C} s\left\{ \frac{\bar{q} - V_\eta^\pi(x, c)}{h_{\mathrm{cert}}} \right\} + \lambda_{\mathrm{mis}} \mathbf{1}\{V_\eta^\pi(x, y) > \bar{q}\}, \tag{3.3}$$

so $0 \leq \ell_{\mathrm{safe}} \leq B_0 = 1 + \lambda_{\mathrm{mis}}$, where $h_{\mathrm{cert}} > 0$ is fixed in advance, $\lambda_{\mathrm{mis}} \geq 0$ penalizes exclusion of the true label, and V_η^π denotes the score obtained under route π. The first term measures normalized smooth set size, while the second prevents apparent efficiency gains obtained by excluding the true label. This loss is used only for certification and never as the final conformal score. For certification point i and active expert e, define the deletion contrast $D_{ie} = \ell_{\mathrm{safe}}(X_i, Y_i; \pi_i^{+e}) - \ell_{\mathrm{safe}}(X_i, Y_i; \pi_i^{-e}) \in [-B_0, B_0]$, where π_i^{+e} is the proposed pre-certification route containing e, while π_i^{-e} deletes e, renormalizes the surviving weights, and falls back to target-only if no expert remains. For expert-cell pair (e, j) with count $N_{ej} \geq n_{\min} \geq 2$, sample mean $\bar{D}_{ej}$, sample variance $\hat{s}_{ej}^2$, and $L_\delta = \log(2MJ/\delta)$, we use

$$U_{ej} = \bar{D}_{ej} + \sqrt{2\hat{s}_{ej}^2 L_\delta / N_{ej}} + \frac{14 B_0 L_\delta}{3(N_{ej} - 1)}. \tag{3.4}$$

The deployed route retains e on G_j only if $U_{ej} \leq 0$; pairs with $N_{ej} < n_{\min}$ are uncertified and rejected. This certifies a cell-average effect, not pointwise safety. The finite partition strikes a practical

balance; it is less flexible than a neural critic, but each retained expert-cell decision has an auditable one-sided bound. The certification sample is independent of data used to fit the deployed predictor and partition, so the empirical-Bernstein calculation applies after conditioning on fitted objects.

Algorithm 1 ROUTECP: learning, certification, and calibration

Require: Expert banks; $\mathcal{D}_0^{\mathrm{tr}}, \mathcal{D}_0^{\mathrm{sel}}, \mathcal{D}_0^{\mathrm{cal}}$; grid Λ; levels α, δ.
1: Split $\mathcal{D}_0^{\mathrm{sel}}$ in advance into $\mathcal{D}_0^{\mathrm{val}}$ and $\mathcal{D}_0^{\mathrm{cert}}$.
2: For each $\lambda \in \Lambda$, fit transports, router, predictor, and score parameters on $\mathcal{D}_0^{\mathrm{tr}}$ by (3.1).
3: On $\mathcal{D}_0^{\mathrm{val}}$, evaluate ordinary foldwise conformal sets, coverage, set size, and activated-expert FLOPs.
4: Select the smallest-set feasible candidate and refit it once on $\mathcal{D}_0^{\mathrm{tr}} \cup \mathcal{D}_0^{\mathrm{val}}$.
5: Construct and freeze $\mathcal{G}$ and certificate reference quantities without using $\mathcal{D}_0^{\mathrm{cert}}$ or $\mathcal{D}_0^{\mathrm{cal}}$.
6: On $\mathcal{D}_0^{\mathrm{cert}}$, compute deletion contrasts and U_{ej} in (3.4); delete uncertified experts, renormalize survivors, and fall back to target-only if none remains.
7: Freeze the predictor, score, partition, and certified route; access $\mathcal{D}_0^{\mathrm{cal}}$ for the first time and compute $\hat{C}_\alpha(x)$.
8: **return** $\hat{C}_\alpha(x)$, certified route, fallback indicator, and route disagreement.

After freezing the predictor, route, partition, certificate, and $\hat{\eta}$, access $\mathcal{D}_0^{\mathrm{cal}}$ for the first time. With calibration scores $v_{(1)} \le \cdots \le v_{(n_{\mathrm{cal}})}$ and $k_\alpha = \lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil$, set $\hat{q}_{1-\alpha} = v_{(k_\alpha)}$ (or $+\infty$ if $k_\alpha = n_{\mathrm{cal}}+1$) and $\hat{C}_\alpha(x) = \{y \in [C] : V_{\hat{\eta}}(x, y; \hat{p}_{\mathrm{cert}}) \le \hat{q}_{1-\alpha}\}$. Algorithm 1 records the complete data flow; the ordering is part of the method, not merely an implementation choice. The four target roles should not be interchanged; training learns representations and routes, validation chooses among candidates, certification removes unsupported expert-cell contributions, and calibration alone determines the final rank threshold. This is the source of the firewall result in Section 4.

4 THEORETICAL PROPERTIES

For a predictor f with nonconformity score V, define $q_\alpha(f, V) = \inf\{t : \mathbb{P}_{P_0}(V(X,Y;f) \le t) \ge 1-\alpha\}$, and let $\mathcal{R}_\alpha(f,V) = \mathbb{E}_{P_0}|\{c \in [C] : V(X,c;f) \le q_\alpha(f,V)\}|$. For candidate λ, $\mathcal{H}_\lambda$ denotes the induced joint predictor-score class; this avoids an oracle comparison that silently changes the conformity score while holding only the predictor fixed. Assume the router-logit class $\mathcal{R}$ obeys $\log N(\epsilon, \mathcal{R}, L_2(Q)) \le V_r \log(A/\epsilon), 0 < \epsilon \le 1$, for every empirical Q and constants $A \ge e$ and $V_r \ge 1$, and impose bounded-complexity source/transport classes. Efficiency additionally uses (i) a positive, bounded target score density near its oracle quantile and (ii) the set-boundary margin $\sum_c \mathbb{P}\{|V(X,c) - q_\alpha| \le t\} \le L_{\mathrm{set}} t$ for some $L_{\mathrm{set}} < \infty$ and all sufficiently small $t > 0$. They localize efficiency sensitivity to score-boundary perturbations. These assumptions are not needed for validity. Let $\mathcal{E}^* = \{e : \mathbb{P}_{P_0}(a_e^*(X) \ne 0) > 0\}$ and $\varepsilon_T = \sum_{e \in \mathcal{E}^*} \omega_e \|\hat{T}_e \hat{h}_e - T_e^* h_e^*\|_{L_2(P_0)}$, $\omega_e \ge 0$, $\sum_{e \in \mathcal{E}^*} \omega_e = 1$, where h_e^* and T_e^* are the population expert and transport maps and ω_e records the oracle usage weight of expert e.

Theorem 4.1. *Let $\hat{\lambda} \in \Lambda$ denote the candidate selected on $\mathcal{D}_0^{\mathrm{val}}$, and let $(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}}) \in \mathcal{H}_{\hat{\lambda}}$ denote the predictor-score pair obtained after training, selection, and refitting steps, but before the finite-partition certificate. The finite grid Λ is fixed before $\mathcal{D}_0^{\mathrm{val}}$ is inspected, and write $q_{\max} = \max_{\lambda \in \Lambda} q_\lambda$. Under the conditions above and the assumptions in Appendix A, with probability at least $1-\delta$,*

$$R_\alpha(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}}) - \inf_{\lambda \in \Lambda} \inf_{(f,V) \in \mathcal{H}_\lambda} R_\alpha(f,V) \le C \Bigg[\underbrace{\sqrt{\frac{q_{\max}\log(eM/q_{\max}) + V_r \log(An_{\mathrm{tr}}) + \log(|\Lambda|/\delta)}{n_{\mathrm{tr}}}}}_{\textit{local-route estimation}}$$

$$+ \underbrace{\sqrt{\frac{\log(|\Lambda|/\delta)}{n_{\mathrm{val}}}}}_{\textit{candidate selection}} + \underbrace{\varepsilon_T}_{\textit{source/transport error}} + \underbrace{\varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{smooth}}}_{\textit{optimization/surrogate error}} \Bigg].$$

If the candidate classes are misspecified relative to the unrestricted population target oracle $f_0^\star$, the right-hand side additionally contains $\mathcal{A}(\Pi_{q_{\hat{\lambda}}})$.

The terms measure uniform sparse-route estimation over the candidate grid, validation selection, source/transport error, optimization error, and surrogate approximation. The use of $q_{\max}$ reflects the simultaneous training bound required before $\hat{\lambda}$ is selected; it does not imply that every deployed route uses $q_{\max}$ experts. Appendix A shows that replacing the population quantile by the calibration order statistic contributes $O_{\mathbb{P}}(n_{\mathrm{cal}}^{-1/2})$ to expected set size under the same local score-regularity conditions.

Theorem 4.2. *Let $\mathcal{R}^*_\alpha$ be the conformal risk of the regionwise population oracle in the family. There exists a two-source, two-region partial-sharing family where each source contains one region-matched expert and one shortcut expert, with each region having mass at least $\rho > 0$ and shortcut harm at least $\kappa > 0$. Every possibly data-dependent rule that accepts or rejects each source as a whole satisfies $\inf_{\hat{s}} \sup_{P \in \mathcal{P}_{\mathrm{part}}} \mathbb{E}_P\{\mathcal{R}_\alpha(\hat{s}) - \mathcal{R}^*_\alpha\} \geq c_0(\rho, \kappa) > 0$, whereas a local router containing the two-region oracle has regret $O(\sqrt{(V_r + \log M)/n_{\mathrm{tr}}} + \sqrt{\log|\Lambda|/n_{\mathrm{val}}} + \varepsilon_T + \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{smooth}})$.*

A whole-source rule makes one accept/reject decision per source, shared by all experts and target inputs, while otherwise depending arbitrarily on the observed data. This restriction matters; whole-source decisions may be structurally unable to attain the local oracle, whereas input-dependent routing adapts transfer to target regions. The explicit construction is given in Appendix A.

For certificate cell G_j, define $\Delta_{ej} = \mathbb{E}D_e(X,Y) \mid X \in G_j, e \in I_q(X)$ as expert e's population cell-average deletion effect in G_j, conditional on its proposal by the pre-certification router.

Proposition 4.3. *Conditional on the predictor, pre-certification router, partition $\mathcal{G}$, and certificate reference quantities fitted without $\mathcal{D}_0^{\mathrm{cert}}$, if the certification observations are independent and identically distributed (i.i.d.) and $0 \leq \ell_{\mathrm{safe}} \leq B_0$, then with probability at least $1-\delta$, $\Delta_{ej} \leq U_{ej}$ simultaneously for every expert-cell pair (e,j) satisfying $N_{ej} \geq n_{\min}$. Consequently, $\mathbb{P}\{\exists(e,j) : U_{ej} \leq 0,\ \Delta_{ej} > 0\} \leq \delta$. Separately, conditional on $\mathcal{F}_{\mathrm{fit}}$, if the target calibration observations and (X_*, Y_*) are exchangeable, then $\mathbb{P}\{Y_* \in \hat{C}_\alpha(X_*) \mid \mathcal{F}_{\mathrm{fit}}\} \geq 1 - \alpha$.*

The first statement is cell-average instead of pointwise. Define $\Delta_e(x) = \mathbb{E}\{D_e(X,Y) \mid X = x,\ e \in I_q(X)\}$ and $\mathrm{osc}_j = \sup_{x,x' \in G_j; e \in \mathcal{E}} |\Delta_e(x) - \Delta_e(x')|$, whenever the conditional expectation exists. Writing $[u]_+ = \max(u, 0)$, on the simultaneous event,

$$\mathbb{E}_{P_0}\left[\sum_{e \text{ retained}; e \in I_q(X)} [\Delta_e(X)]_+\right] \leq q \sum_{j=1}^{J} P_0(G_j)\,\mathrm{osc}_j\,. \tag{4.1}$$

The factor q follows because at most q experts can be active at any input. Thus finer cells improve localization but reduce N_{ej} and widen the certificate. By contrast, the coverage guarantee requires none of the partial-sharing, entropy, alignment, optimization, or certification assumptions; after all fitted objects are frozen, it follows solely from calibration-test exchangeability.

5 Numerical Studies

We evaluate the proposed framework under partial transfer, focusing on conformal efficiency, harmful-route protection, and target-only fallback when transfer is unsupported. All procedures target 90% marginal coverage; hyperparameter grids, splits, seeds, and baseline budgets are fixed before test outcomes are inspected. Baselines include target-only learning, dense source transfer, whole-source selection, global expert gates, weighted conformal prediction (Tibshirani et al., 2019), multi-source conformal prediction (Liu et al., 2024), transfer conformal prediction (Zhang et al., 2026), and a parameter-matched sparse mixture-of-experts router trained without the conformal-efficiency objective.

5.1 Monte Carlo Simulations

We consider a five-class problem ($C = 5$) with four observable target regions, so the oracle expert support is known for evaluation but never supplied to the router. We draw $X \sim \mathcal{N}(0, I_{20})$ and define $G_r = \{x : b_{r-1} < u^\top x \leq b_r\}$, where $u = (1,1,1,1,0,\ldots,0)^\top/2$ and the cut points are population quartiles, so $P_0(G_r) = 1/4$. Four sources expose four experts each ($M = 16$). Each region has a different two-expert oracle support; six additional experts are neutral there and the remaining eight contain region-dependent shortcuts. The target logit satisfies $f_0^*(x) = f^*_{\mathrm{private}}(x) + \sum_{e=1}^{16} a_e^*(x) f_e^*(x)$ with $\|a^*(x)\|_0 \leq 2$, and $Y \mid X = x \sim \mathrm{Categorical}\{\mathrm{softmax}(f_0^*(x))\}$. Source representations are independently rotated, rescaled, and perturbed before transport, so successful transfer requires both alignment and support recovery. Appendix C gives the complete DGP, region-specific supports, expert functions, shortcut construction, and representation transformations. Unless varied, $n_s = 4000$, $n_{\mathrm{tr}} = 800$, $n_{\mathrm{val}} = 400$, $n_{\mathrm{cert}} = 400$, $n_{\mathrm{cal}} = 800$, and $n_{\mathrm{test}} = 5000$; the routing budget is $q = 2$, shortcut strength is $\zeta = .8$, and nominal coverage is 90%. We use 500 replications for performance comparisons and 1000 for certificate failure. The default certificate uses $J = 8$ cells, $n_{\min} = 20$, and $\delta = .05$.

Table 1: **Main Monte Carlo comparison.** Excess risk is population conformal risk relative to the oracle route. Route F1 compares estimated and oracle input-specific supports; NT is the fraction of replications in which transfer has larger population conformal risk than the paired target-only fit; Experts is the mean number of source experts evaluated after certification. Standard errors are in parentheses.

Method	Cov.	Size↓	Excess↓	Route F1↑	NT↓	Experts↓
Target only	.903(.002)	1.94(.02)	.236(.010)	–	–	0
Dense transfer	.900(.002)	2.10(.03)	.397(.014)	–	.418	16.0
Whole-source	.901(.002)	1.92(.02)	.219(.011)	.48(.02)	.196	7.6
Weighted CP	.904(.002)	1.87(.02)	.169(.009)	–	.151	–
Multi-source CP	.903(.002)	1.82(.02)	.123(.008)	–	.124	–
Transfer CP	.902(.002)	1.80(.02)	.109(.008)	–	.112	–
Global expert gates	.902(.002)	1.75(.02)	.069(.007)	.72(.02)	.091	3.8
Parameter-matched router	.901(.002)	1.69(.02)	.050(.006)	.79(.01)	.063	2.00
RouteCP w/o certificate	.901(.002)	1.57(.02)	.017(.004)	.85(.01)	.067	2.00
RouteCP	.903(.002)	1.61(.02)	.027(.005)	.83(.01)	.014	1.73
Oracle route	.902(.002)	1.54(.01)	0	1.00	0	2.00

Table 1 summarizes the default partial-transfer regime. Dense transfer can underperform target-only learning because shortcuts are imported with useful experts. Whole-source selection helps but cannot separate useful and harmful components within a source. The parameter-matched router shows that routing capacity alone does not explain the gain. Certification increases mean set size by about 0.04 labels while reducing paired negative-transfer frequency from .067 to .014.

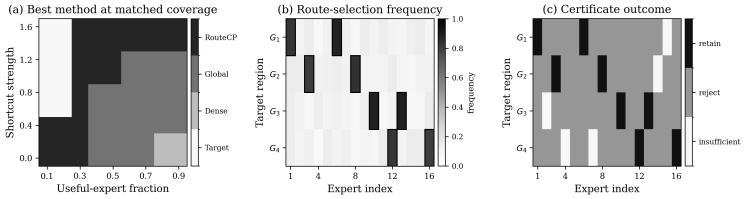


Figure 2: **Partial transfer, route recovery, and safety.** (a) Smallest mean set size among methods attaining at least 89.5% empirical coverage; gray levels denote methods. (b) Region-specific expert-selection frequencies; the color bar gives selection probability and boxes mark oracle support. (c) Certificate decisions; the legend distinguishes insufficient support, rejection, and retention.

Figure 2 identifies where local routing is useful. Dense sharing is competitive when most experts are transferable and shortcuts are weak, whereas target-only prediction dominates when transferable structure is nearly absent. The largest advantage occurs between these extremes, when useful and harmful computations coexist within a source and the appropriate support varies with X.

Table 2: **Mechanism diagnostics.** The panels examine structural separation, finite-partition safety, and component ablations. FA is the familywise false-acceptance event in Proposition 4.3; FR is target-only fallback frequency.

Panel A: structural separation, excess conformal risk

n_{tr}	Target only	Whole-source	Global gates	RouteCP
200	.287(.018)	.215(.016)	.146(.013)	.101(.011)
400	.253(.015)	.193(.014)	.111(.010)	.064(.008)
800	.234(.012)	.179(.011)	.083(.008)	.036(.006)
1600	.220(.010)	.171(.010)	.064(.007)	.021(.004)
3200	.212(.009)	.168(.009)	.054(.006)	.012(.003)

Panel B: certificate resolution at nominal $\delta = .05$

Cells J	Median N_{ej}	FA [95% CI]↓	NT↓	Size↓	FR
4	82	.032 [.023,.045]	.029	1.68	.109
8	41	.021 [.014,.032]	.014	**1.61**	.147
16	21	.015 [.009,.025]	.010	1.64	.226
32	10	.013 [.008,.022]	.009	1.74	.381

Panel C: ablations under controlled partial sharing

Variant	Coverage	Size↓	Route F1↑	NT↓	FR
Full RouteCP	.903	1.61	.83	.014	.147
No conformal objective	.902	1.72	.79	.026	.151
No alignment	.901	1.81	.67	.052	.211
No disagreement factor	.902	1.65	.83	.020	.141
No certificate	.901	**1.57**	**.85**	.067	0
Global router	.902	1.75	.72	.091	.082

For the certificate-resolution experiment, each J defines a separate partition rule fixed before its certification sample is inspected; J is not selected using the reported false-acceptance outcomes. Table 2

connects the numerical evidence to the theory. Panel A targets Theorem 4.2; the whole-source gap remains bounded away from zero, whereas local-routing regret decreases with n_{tr}. Panel B evaluates Proposition 4.3 through its exact familywise false-acceptance event instead of using negative transfer as a proxy. The intermediate partition gives the smallest sets, illustrating the localization-support tradeoff. Panel C shows the division of labor; alignment supports expert reuse, the conformal objective targets set efficiency, disagreement modifies scores in uncertain regions, and certification supplies a conservative post-training filter. None is required for the final exchangeable-rank coverage guarantee.

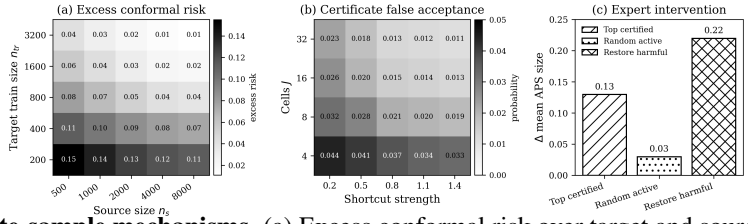


Figure 3: **Finite-sample mechanisms.** (a) Excess conformal risk over target and source sample sizes. (b) Empirical familywise false acceptance over certificate resolution and shortcut strength. (c) Change in mean APS size under expert deletion and restoration interventions.

Figure 3 supports the finite-sample interpretation of Theorem 4.1; increasing n_{tr} chiefly improves route discovery, whereas larger source samples help when expert estimation is a leading error component. The intervention panel provides an operational check on route mass; deleting the highest-mass certified expert increases set size significantly more than deleting a random active expert, while restoring the most harmful rejected expert sharply increases negative transfer.

5.2 Empirical Applications

We use PACS, OfficeHome, Camelyon17-WILDS, and frozen CLIP encoders on PACS/OfficeHome. PACS and OfficeHome use each domain as target and the remaining domains as sources. Camelyon17 follows hospital-based WILDS environments but uses a prespecified small labeled target subset for adaptation, certification, and calibration; the remaining target-hospital images are test data. Residual stages, adapters, grouped attention blocks, and LoRA updates form the expert bank. MCCD is the maximum classwise coverage deviation and is reported only as a diagnostic, not as a guaranteed property.

Table 3: **Empirical performance and target-domain behavior at 90% marginal coverage.** Panel A averages over target domains, splits, and seeds; Panel B reports benchmark-level behavior of RouteCP. Additional benchmark- and domain-level diagnostics are reported in Appendix F.

Panel A: aggregate baselines

Method	Acc.↑	Cov.	APS size↓	MCCD↓	Rel. FLOPs↓
Target only	78.4(.3)	.902(.002)	2.31(.03)	.071(.004)	1.00
Full target fine-tuning	79.6(.3)	.901(.002)	2.24(.03)	.069(.004)	1.00
Dense source transfer	80.0(.3)	.899(.002)	2.23(.03)	.072(.004)	2.79
Whole-source selection	80.5(.3)	.902(.002)	2.11(.03)	.063(.003)	1.91
AdapterFusion	80.8(.3)	.901(.002)	2.04(.03)	.060(.003)	1.83
Weighted CP	80.2(.3)	.904(.002)	2.07(.03)	.061(.003)	1.39
Multi-source CP	80.7(.3)	.903(.002)	2.01(.03)	.058(.003)	1.70
Transfer CP	80.9(.3)	.902(.002)	1.98(.03)	.057(.003)	1.74
Global expert gates	81.3(.3)	.902(.002)	1.96(.02)	.055(.003)	1.61
Parameter-matched router	81.6(.3)	.901(.002)	1.88(.02)	.052(.003)	1.48
RouteCP w/o certificate	82.1(.2)	.900(.002)	1.75(.02)	.049(.003)	1.47
RouteCP	82.0(.2)	.903(.002)	1.79(.02)	.045(.003)	1.44

Panel B: per-benchmark target-only versus certified RouteCP

Benchmark	TO acc.	RouteCP acc.	TO size	RouteCP size	FR
PACS	84.2	87.1	1.81	1.49	.103
OfficeHome	71.8	75.1	2.70	2.07	.161
Camelyon17	76.4	78.0	2.55	2.18	.244
CLIP-PACS	87.6	90.1	1.69	1.35	.077
CLIP-OfficeHome	80.9	85.0	2.29	1.73	.091

Table 3 shows that source-level baselines improve on target-only prediction but leave a gap to module-level routing. The parameter-matched router is stronger than global gating but less efficient than the conformal-efficiency-trained router, suggesting that the gain is not merely model capacity. Certification increases mean set size from 1.75 to 1.79 but improves MCCD and suppresses harmful activations. The largest gains appear for frozen vision-language experts, while Camelyon17 has the largest fallback rate, consistent with stronger hospital shift.

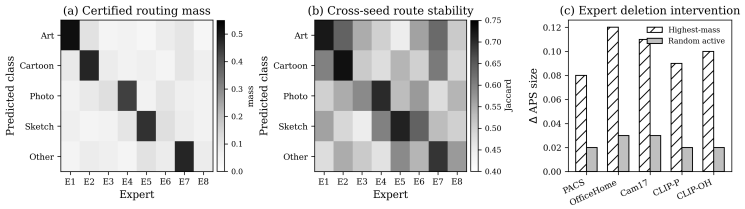


Figure 4: **Routing in real pretrained models.** Left: mean certified routing mass across target inputs. Middle: cross-seed route stability under independent optimization runs. Right: deletion intervention comparing the highest-mass certified expert with a randomly selected active expert.

The routing atlas in Figure 4 checks that the learned supports are not merely global domain preferences. Early modules are often reused broadly, whereas later adapters and LoRA experts are more class- and input-specific. Certified supports have mean cross-seed Jaccard agreement about 0.73; deleting the highest-mass certified expert increases APS set size by about 0.10 labels, compared with 0.024 for a random active expert. This supports interpreting route mass as operationally meaningful.

Table 4: **Empirical robustness diagnostics.** Low-resource cells report APS size; stress tests report ROUTECP minus target-only APS size, fallback, and targeted harmful-route retention.

Panel A: low-resource target adaptation					
Method	25% train	50% train	100% train	–	–
Target only	2.79	2.51	2.31	–	–
Whole-source	2.48	2.28	2.11	–	–
Global gates	2.29	2.10	1.96	–	–
Parameter-matched router	2.16	1.99	1.88	–	–
ROUTECP	1.99	1.87	1.79	–	–

Panel B: transfer stress tests					
Condition	Coverage	ΔSize	Fallback	Harmful activation	Note
No artificial stress	.903	-.52	.14	.018	useful transfer
Permuted source labels	.902	+.02	.93	.011	mostly reject
Severe misalignment	.904	+.04	.64	.026	fallback
Reversed shortcut	.901	-.08	.47	.034	certified
Reversed shortcut, no cert.	.900	-.15	0	.151	unsafe gain

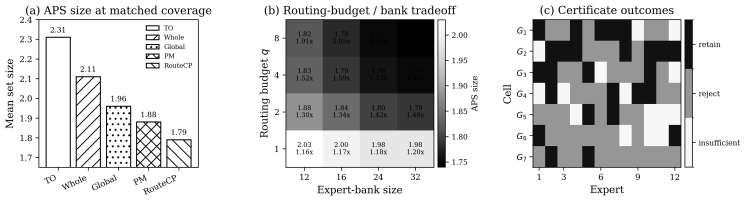


Figure 5: **Efficiency, computation, and certification.** Left: matched-coverage APS size. Middle: routing-budget/expert-bank tradeoff; each cell reports APS size and relative FLOPs. Right: certificate outcomes for insufficient support, rejection, and retention across expert-cell pairs.

Table 4 and Figure 5 show that the empirical gains are consistent with the intended operating regime of ROUTECP. Source reuse is most beneficial when target fitting labels are scarce, while the independent calibration split remains necessary for the final coverage guarantee. At the selected routing budgets, ROUTECP evaluates about 2.2 source experts per input from banks of 12-32 candidates, yielding mean inference cost of about $1.44\times$ target-only compared with $2.79\times$ under dense evaluation, without requiring dense source pre-evaluation. When source labels are permuted or transport alignment is severely misspecified, the certificate rejects most transferred routes and the deployed predictor falls back towards target-only behavior. Under a reversed shortcut, certification gives up part of the apparent set-size improvement while significantly reducing harmful expert activation.

6 CONCLUDING REMARKS

ROUTECP combines input-dependent expert reuse, held-out safety certification, and untouched conformal calibration; its theory and experiments characterize efficiency gains and conservative fallback when transfer is unsupported. Finite-sample marginal coverage follows from calibration-test exchangeability, whereas efficiency and safety depend on partial-sharing and certification conditions. The finite-partition certificate controls cell-average harmful routing instead of pointwise behavior, and richer transports, adaptive partitions, or nonexchangeable targets require additional assumptions.

AI USE STATEMENT

Generative AI tools were used only to assist with writing and language editing. They were not used to generate research ideas or theoretical results. All AI-assisted text was reviewed and revised by the authors for accuracy, clarity, and appropriate attribution. The authors take full responsibility for the final content of this work, including all claims and text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed assumptions and complete proofs of the theoretical results in the appendix. The appendix also specifies the simulation designs, data processing, model architectures, training and tuning procedures, baselines, and additional experimental results needed to reproduce the empirical study. To preserve anonymity during double-blind review, all code used to generate the experiments, tables, and figures will be made publicly available upon completion of the peer-review process.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from XXX. The authors also thank seminar and conference participants at XXX for helpful comments and discussions.

REFERENCES

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494-591.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2024). Conformal risk control. *International Conference on Learning Representations*.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. (2021). Uncertainty sets for image classifiers using conformal prediction. *International Conference on Learning Representations*.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 1-34.

Einbinder, B.-S., Romano, Y., Sesia, M., and Zhou, Y. (2022). Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 35.

Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1-39.

Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. *International Conference on Learning Representations*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning*, 97, 2790-2799.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Kasa, K., Zhang, Z., Yang, H., and Taylor, G. W. (2025). Adapting prediction sets to distribution shifts without labels. *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, 286, 1990-2010.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of the 38th International Conference on Machine Learning*, 139, 5637-5664.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.

Liu, Y., Levis, A., Normand, S.-L., and Han, L. (2024). Multi-source conformal inference under distribution shift. *Proceedings of the 41st International Conference on Machine Learning*, 235, 31344-31382.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2018). Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 31.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. (2019). Moment matching for multi-source domain adaptation. *IEEE International Conference on Computer Vision*, 1406-1415.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. (2021). AdapterFusion: Non-destructive task composition for transfer learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 487-503.

Plassier, V., Kotelevskii, N., Rubashevskii, A., Noskov, F., Velikanov, M., Fishkov, A., Horvath, S., Takac, M., Moulines, E., and Panov, M. (2024). Efficient conformal prediction under data heterogeneity. *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 238, 4879-4887.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A. S., Keysers, D., and Houlsby, N. (2021). Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34.

Romano, Y., Sesia, M., and Candès, E. J. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12), 371-421.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations*.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer Verlag, Berlin.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. (2019). Characterizing and avoiding negative transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11293-11302.

Xu, R., Chen, C., Sun, Y., Venkitasubramaniam, P., and Xie, S. (2025). Wasserstein-regularized conformal prediction under general distribution shift. *International Conference on Learning Representations*.

Zhang, C., Li, T., Xie, J., Kong, L., and Jiang, B. (2026). Transfer conformal predictive inference for regression. *Journal of Machine Learning Research*, 27(90), 1-68.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V. Y., Dai, A. M., Chen, Z., Le, Q. V., and Laudon, J. (2022). Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35.

APPENDIX

A THEORETICAL DETAILS AND PROOFS

This appendix gives the assumptions and proofs supporting Section 4. The statistical arguments follow the data-access order in Algorithm 1; source and target training data are used for fitting, validation data are used for candidate selection and the single refit, certification data are used only after that refit, and calibration data are opened only after the deployed score has been frozen. The final coverage statement is the usual exchangeable-rank argument for split conformal prediction (Vovk et al., 2005; Lei et al., 2018; Angelopoulos and Bates, 2023). The efficiency analysis is separate; it concerns the learned local router before the safety filter, while Proposition 4.3 controls the subsequent certification stage. This distinction is needed because a conservative certificate can delete a useful but weakly supported expert and therefore need not preserve the pre-certification oracle risk.

A.1 FORMAL CLASSES, REGULARITY CONDITIONS, AND AUXILIARY BOUNDS

We write $\mathcal{D}_0^{\mathrm{sel}} = \mathcal{D}_0^{\mathrm{val}} \dot{\cup} \mathcal{D}_0^{\mathrm{cert}}$. Let $\mathcal{F}_{\mathrm{tr}}$ contain the source samples, $\mathcal{D}_0^{\mathrm{tr}}$, the prespecified expert decomposition, the candidate grid Λ, and all fitting randomness used before validation. Let $\mathcal{F}_{\mathrm{val}}$ additionally contain $\mathcal{D}_0^{\mathrm{val}}$, the selected candidate $\hat{\lambda}$, and the predictor obtained by the single refit on $\mathcal{D}_0^{\mathrm{tr}} \cup \mathcal{D}_0^{\mathrm{val}}$. Let $\mathcal{F}_{\mathrm{pre}}$ additionally contain the finite partition $\mathcal{G}$ and all certificate reference quantities constructed without $\mathcal{D}_0^{\mathrm{cert}}$ or $\mathcal{D}_0^{\mathrm{cal}}$, and let $\mathcal{F}_{\mathrm{fit}}$ additionally contain the certification sample and the resulting frozen certified route. Thus $\mathcal{D}_0^{\mathrm{cal}}$ is not measurable with respect to $\mathcal{F}_{\mathrm{fit}}$. For a predictor-score pair (f, V), define

$$q_\alpha(f, V) = \inf\{t : P_0(V(X, Y; f) \le t) \ge 1 - \alpha\},$$

$$\mathcal{R}_\alpha(f, V) = \mathbb{E}_{P_0} \sum_{c=1}^{C} \mathbf{1}\{V(X, c; f) \le q_\alpha(f, V)\}.$$

For candidate λ, $\mathcal{H}_\lambda$ is the joint class of predictor-score pairs generated by the admissible target backbone, affine transports, prediction head, router, transfer amplitude, and score parameters under that candidate. The oracle comparison is therefore joint in (f, V), instead of fixing the predictor while varying the conformity score. The following assumptions specify the approximation, complexity, and regularity conditions under which these candidate classes are analyzed.

Assumption 1. *The target training, validation, certification, and calibration samples are mutually independent i.i.d. samples from P_0 for the efficiency and certification results. For the final coverage statement only, the weaker requirement is used; conditional on $\mathcal{F}_{\mathrm{fit}}$, the n_{cal} calibration observations and the future target observation (X_*, Y_*) are exchangeable. The source samples are independent of the target splits but may follow arbitrary source distributions P_s.*

The separation in Assumption 1 mirrors the deployed procedure. Independence of training, validation, and certification allows the learning and safety arguments to condition on previously fitted objects, whereas the final coverage statement requires only calibration-test exchangeability. Consequently, source-target similarity is never used to justify coverage; it enters only through the efficiency model below. This distinction isolates predictive-transfer assumptions from the distribution-free validity mechanism, preserving coverage even under severe source-domain mismatch.

Assumption 2. *There are finite constants $B_h, B_g, B_V, \gamma_{\max}$ such that all aligned expert outputs, fitted logits, and conformity scores are uniformly bounded by the corresponding constants. The target-backbone, affine-transport, and prediction-head classes have bounded empirical metric entropy at the scales used below. For every empirical measure Q and $0 < \epsilon \le 1$, the router-logit class satisfies*

$$\log N(\epsilon, \mathcal{R}, L_2(Q)) \le V_r \log(A/\epsilon), \quad A \ge e,$$

and the class of input-dependent top-q support maps induced by $\mathcal{R}$ has empirical growth

$$\log \Pi_{q,\mathcal{R}}(n) \le C_I\{q \log(eM/q) + V_r \log(An)\}. \tag{A.1}$$

The temperature satisfies $\tau \ge \tau_{\min} > 0$ over the analyzed candidate grid. Condition (A.1) is automatic for the finite-dimensional router families used in the experiments after replacing V_r by a constant multiple of their pseudo-dimension; it is stated explicitly because an L_2 covering bound for logits alone does not control the discontinuous input-dependent top-q support map.

The explicit support-growth condition is important. A proof that merely unions over the $\binom{M}{q}$ static supports would be insufficient because one router may use different supports at different target inputs. Equation (A.1) controls the number of support labelings of the observed sample and therefore covers the actual input-varying routing class without assuming a fixed support pattern.

Assumption 3. *The centered target logit admits*

$$f_0^*(x) = f^*_{\text{private}}(x) + \sum_{e\in\mathcal{E}} a_e^*(x) f_e^*(x) + r_q(x), \quad \|a^*(x)\|_0 \le q, \quad \|a^*(x)\|_\infty \le A_{\max},$$

where r_q is absorbed by the approximation error $\mathcal{A}(\Pi_q)$. With $\mathcal{E}^* = \{e : P_0(a_e^*(X) \neq 0) > 0\}$ and nonnegative ω_e summing to one,

$$\varepsilon_T = \sum_{e\in\mathcal{E}^*} \omega_e \|\hat{T}_e\hat{h}_e - T_e^* h_e^*\|_{L_2(P_0)} < \infty$$

with the probability stated in the theorem.

Assumption 3 formalizes the intended notion of partial transfer. It permits the active support and coefficient magnitudes to vary with the target input, while preventing the population representation from becoming vacuous by restricting each reusable contribution to a source-induced class and imposing local sparsity. The residual r_q is not hidden estimation error; it is the approximation component already represented by $\mathcal{A}(\Pi_q)$ in the main theorem. The next lemma isolates how source-estimation and transport errors propagate into this approximation framework.

Lemma A.1. *If T_e^* is L_T-Lipschitz on the support of the relevant representations, then*

$$\varepsilon_T \le \sum_{e\in\mathcal{E}^*} \omega_e \left[L_T \|\hat{h}_e - h_e^*\|_{L_2(P_0)} + \|(\hat{T}_e - T_e^*)\hat{h}_e\|_{L_2(P_0)} \right]. \tag{A.2}$$

Hence, if a finite-dimensional source expert satisfies

$$\|\hat{h}_e - h_e^*\|_{L_2(P_0)} = O_{\mathbb{P}}\left(\sqrt{\frac{d_e + \log(M/\delta)}{n_{s(e)}}} \right),$$

then its contribution through source-expert estimation has the same order, while the second term records residual alignment error or transport misspecification.

Proof. We fix $e \in \mathcal{E}^*$. Adding and subtracting $T_e^*\hat{h}_e$ gives the exact decomposition

$$\hat{T}_e\hat{h}_e - T_e^* h_e^* = \{\hat{T}_e\hat{h}_e - T_e^*\hat{h}_e\} + \{T_e^*\hat{h}_e - T_e^* h_e^*\} = (\hat{T}_e - T_e^*)\hat{h}_e + T_e^*(\hat{h}_e) - T_e^*(h_e^*).$$

Taking the $L_2(P_0)$ norm and applying Minkowski's inequality therefore yields

$$\|\hat{T}_e\hat{h}_e - T_e^* h_e^*\|_{L_2(P_0)} \le \|(\hat{T}_e - T_e^*)\hat{h}_e\|_{L_2(P_0)} + \|T_e^*(\hat{h}_e) - T_e^*(h_e^*)\|_{L_2(P_0)}.$$

By the assumed Lipschitz property, for P_0-almost every x, we have

$$\|T_e^*\{\hat{h}_e(x)\} - T_e^*\{h_e^*(x)\}\|_2 \le L_T \|\hat{h}_e(x) - h_e^*(x)\|_2.$$

Squaring both sides, integrating with respect to P_0, and taking square roots gives

$$\|T_e^*(\hat{h}_e) - T_e^*(h_e^*)\|_{L_2(P_0)} \le L_T \|\hat{h}_e - h_e^*\|_{L_2(P_0)}.$$

Combining the preceding two displays gives, for every $e \in \mathcal{E}^*$,

$$\|\hat{T}_e\hat{h}_e - T_e^* h_e^*\|_{L_2(P_0)} \le L_T \|\hat{h}_e - h_e^*\|_{L_2(P_0)} + \|(\hat{T}_e - T_e^*)\hat{h}_e\|_{L_2(P_0)}.$$

Multiplying by the nonnegative oracle usage weight ω_e and summing over $e \in \mathcal{E}^*$ proves the asserted bound for ε_T in (A.2). For the final statement, substituting

$$\|\hat{h}_e - h_e^*\|_{L_2(P_0)} = O_{\mathbb{P}}\left(\sqrt{(d_e + \log(M/\delta))/n_{s(e)}} \right)$$

into the first term shows that source-expert estimation contributes

$$O_{\mathbb{P}}\left(L_T \sqrt{\frac{d_e + \log(M/\delta)}{n_{s(e)}}} \right)$$

for expert e (with fixed L_T). The remaining quantity $\|(\hat{T}_e - T_e^*)\hat{h}_e\|_{L_2(P_0)}$ is not absorbed into that rate, because it separately captures estimation or misspecification of the transport map. This proves the claim. □

Assumption 4. *For each candidate λ, choose $(f_\lambda^\star, V_\lambda^\star) \in \arg\min_{(f,V)\in\mathcal{H}_\lambda} R_\alpha(f,V)$; when the infimum is not attained, take an $o(1)$ approximate minimizer. Let $\mathcal{L}_\lambda$ and $(f_\lambda^\circ, V_\lambda^\circ)$ be the population smoothed objective and its minimizer. Uniformly over the finite grid, and locally around this minimizer, assume $\|V_\lambda^\circ - V_\lambda^\star\|_{L_2(P_0\times\mathrm{Unif}[C])} \le \varepsilon_{\mathrm{smooth}}$,*

$$\begin{aligned} c_L\|V - V_\lambda^\circ\|^2_{L_2(P_0\times\mathrm{Unif}[C])} &\le \mathcal{L}_\lambda(f,V) - \mathcal{L}_\lambda(f_\lambda^\circ, V_\lambda^\circ), \\ \mathrm{Var}\{\ell_{f,V} - \ell_{f_\lambda^\circ, V_\lambda^\circ}\} &\le C_L\{\mathcal{L}_\lambda(f,V) - \mathcal{L}_\lambda(f_\lambda^\circ, V_\lambda^\circ)\}. \end{aligned} \tag{A.3}$$

The loss class is locally Lipschitz and bounded. On this neighborhood, score distributions obey

$$\sup_t \left|P_0\{V(X,Y)\le t\} - P_0\{V_\lambda^\circ(X,Y)\le t\}\right| \le C_{\mathrm{sp}}\|V - V_\lambda^\circ\|_{L_2(P_0\times\mathrm{Unif}[C])}, \tag{A.3a}$$

with the analogous candidate-label averaged bound. The returned fit is within its prescribed optimization tolerance and induces at most $\varepsilon_{\mathrm{opt}}$ score perturbation in $L_2(P_0\times\mathrm{Unif}[C])$. Here $\varepsilon_{\mathrm{smooth}}$ measures the deterministic score-scale gap between the surrogate and conformal-risk targets, and does not represent statistical estimation or numerical optimization error. The same localized conditions and score-scale bounds hold uniformly for every hypothetical refit on $\mathcal{D}_0^{\mathrm{tr}} \cup \mathcal{D}_0^{\mathrm{val}}$.

The conditions in Assumption 4 are about the statistical target of the nonconvex learner, not a claim that straight-through optimization finds a global optimum. The quadratic/Bernstein condition is the step that converts a localized empirical-process bound into an L_2 score bound at the square-root rate. Without such a condition, small excess surrogate objective would not by itself imply the score perturbation used in the conformal-efficiency argument. Localized empirical-process bounds of this form are standard under entropy and Bernstein conditions (van der Vaart and Wellner, 1996; Wainwright, 2019).

Assumption 5. *For every population comparator used below, the true-label score distribution has a density between $c_- > 0$ and $c_+ < \infty$ on a fixed neighborhood of its $(1-\alpha)$ quantile. Moreover, for all sufficiently small $t > 0$,*

$$\sum_{c=1}^{C} P_0\{|V(X,c) - q_\alpha(f,V)| \le t\} \le L_{\mathrm{set}}\, t. \tag{A.4}$$

The density and boundary conditions in Assumption 5 exclude only pathological configurations in which an arbitrarily small score perturbation moves a nonnegligible fraction of labels across the set boundary. They are local conditions around the operating quantile, not global smoothness assumptions on the full score distribution, and they are invoked only when converting score error into set-size error.

Assumption 6. *Let $\lambda^\star$ minimize $\inf_{(f,V)\in\mathcal{H}_\lambda} \mathcal{R}_\alpha(f,V)$ over the finite candidate grid. Its population activated-expert cost is strictly below the prespecified FLOP budget by a fixed slack, and its population coverage satisfies the validation guard by a fixed slack. The same holds for any $o(1)$ approximate minimizer used when the infimum is not attained. Thus the oracle comparator is not excluded by the selection rule for deterministic reasons.*

This last assumption is needed because the implemented selector searches only among candidates satisfying the validation coverage and compute guards. Without population feasibility of the oracle comparator, no procedure restricted to the feasible set can be compared with the unrestricted infimum appearing in the theorem. The next lemma quantifies the finite-sample error introduced by selecting among these feasible candidates on the validation sample.

Lemma A.2. *Under Assumption 5, if two bounded score functions V and V' satisfy $\|V - V'\|_{L_2(P_0\times\mathrm{Unif}[C])} \le r$ and r is sufficiently small, then*

$$|q_\alpha(V) - q_\alpha(V')| \le C_q r \text{ and } |\mathcal{R}_\alpha(V) - \mathcal{R}_\alpha(V')| \le C_R r, \tag{A.5}$$

where the constants depend only on $c_-, c_+, L_{\mathrm{set}}, B_V$, and C.

Proof. We write $q = q_\alpha(V)$ and $q' = q_\alpha(V')$, and let $F(t) = P_0\{V(X,Y)\le t\}$ and $F'(t) = P_0\{V'(X,Y)\le t\}$ be the corresponding true-label score distribution functions. By the localized score-process condition (A.3a), we have $\sup_t |F'(t) - F(t)| \le C_{\mathrm{sp}} r$. By Assumption 5, there exists $\nu > 0$ such that F has a density f_V satisfying

$$0 < c_- \le f_V(t) \le c_+ < \infty \text{ and } t \in [q-\nu, q+\nu].$$

In particular, F is continuous at q, and hence $F(q) = 1-\alpha$. For every $0 < t \le \nu$,

$$F(q+t) - F(q) = \int_q^{q+t} f_V(u)\,du \geq c_- t,$$

and similarly

$$F(q) - F(q-t) = \int_{q-t}^{q} f_V(u)\,du \geq c_- t.$$

We set $t_r = 2C_{\mathrm{sp}} r / c_-$. For sufficiently small r, $t_r \leq \nu$. Using (A.6), we can get

$$F'(q+t_r) \geq F(q+t_r) - C_{\mathrm{sp}} r \geq 1 - \alpha + C_{\mathrm{sp}} r > 1 - \alpha,$$

whereas

$$F'(q-t_r) \leq F(q-t_r) + C_{\mathrm{sp}} r \leq 1 - \alpha - C_{\mathrm{sp}} r < 1 - \alpha.$$

By the definition of the generalized $(1-\alpha)$ quantile, these two inequalities imply $q - t_r < q' \leq q + t_r$. Consequently,

$$|q' - q| \leq \frac{2C_{\mathrm{sp}}}{c_-} r = C_q r.$$

We next control the induced change in population conformal risk. For $c \in [C]$, we write $V_c(X) = V(X, c)$ and $V'_c(X) = V'(X, c)$. Then

$$\mathcal{R}_\alpha(V) = \sum_{c=1}^{C} P_0\{V_c(X) \leq q\} \text{ and } \mathcal{R}_\alpha(V') = \sum_{c=1}^{C} P_0\{V'_c(X) \leq q'\}.$$

Adding and subtracting $\sum_{c=1}^{C} P_0\{V_c(X) \leq q'\}$ gives

$$\begin{aligned} |\mathcal{R}_\alpha(V') - \mathcal{R}_\alpha(V)| \leq & \sum_{c=1}^{C} |P_0\{V'_c(X) \leq q'\} - P_0\{V_c(X) \leq q'\}| \\ & + \sum_{c=1}^{C} |P_0\{V_c(X) \leq q'\} - P_0\{V_c(X) \leq q\}|\,. \end{aligned}$$

The candidate-label version of the localized score-process condition controls the first term uniformly in the threshold

$$\sum_{c=1}^{C} |P_0\{V'_c(X) \leq t\} - P_0\{V_c(X) \leq t\}| \leq C_1 r,$$

where C_1 depends only on the constants in Assumption 5 and on C.

For the second term, let $\Delta_q = |q' - q|$. For each c, the indicators $\mathbf{1}\{V_c(X) \leq q'\}$ and $\mathbf{1}\{V_c(X) \leq q\}$ can differ only when $|V_c(X) - q| \leq \Delta_q$. Hence, by the set-boundary margin condition equation A.4,

$$\sum_{c=1}^{C} |P_0\{V_c(X) \leq q'\} - P_0\{V_c(X) \leq q\}| \leq \sum_{c=1}^{C} P_0\{|V_c(X) - q| \leq \Delta_q\} \leq L_{\mathrm{set}} \Delta_q \leq L_{\mathrm{set}} C_q r.$$

Combining above equations yields

$$|\mathcal{R}_\alpha(V') - \mathcal{R}_\alpha(V)| \leq \big(C_1 + L_{\mathrm{set}} C_q\big) r = C_R r.$$

The constants can be chosen uniformly over the finite candidate grid because the bounds in Assumptions 4-5 are uniform. This proves the lemma.

□

A.2 Oracle Inequality, Separation, and Certification Proofs

We first prove the oracle inequality for the selected and refitted *pre-certification* router. This is the mathematically appropriate efficiency object; the certificate is a separate safety filter whose finite-sample guarantee is Proposition 4.3. A certificate may reject a genuinely useful expert when its held-out support is insufficient, so an oracle inequality for the post-certification predictor requires an additional certification-power term and is not implied by cell-average nonharmfulness alone. We therefore begin by controlling the estimation error for a fixed candidate before accounting for validation-based model selection, which separates within-candidate learning from finite-grid comparison cleanly.

Lemma A.3. *Under Assumptions 2-4, for a fixed candidate λ with budget q_λ, with probability at least $1-\delta$,*

$$\|\hat{V}_\lambda - V_\lambda^\circ\|_{L_2(P_0\times \mathrm{Unif}[C])} \le C\left[\sqrt{\frac{q_\lambda \log(eM/q_\lambda) + V_r \log(An_{\mathrm{tr}}) + \log(1/\delta)}{n_{\mathrm{tr}}}} + \varepsilon_T + \varepsilon_{\mathrm{opt}}\right]. \tag{A.6}$$

Proof. We fix a candidate λ and write $q = q_\lambda$. Through the proof, condition on the source samples and on the fitted source experts and transports on the event defining ε_T. Conditional on these quantities, the remaining randomness is that of the n_{tr} target training observations. Let $\mathcal{V}_\lambda$ be the score class induced by candidate λ, and let $\mathcal{L}_\lambda(V)$ and $\hat{\mathcal{L}}_\lambda(V)$ denote its population and empirical training objectives, respectively. We write V_λ° for the population minimizer within the candidate class and first consider an exact empirical minimizer $\tilde{V}_\lambda$; optimization and source-transport errors will be added at the end.

We begin with the complexity of the routed class. For a target sample $x_1, \dots, x_{n_{\mathrm{tr}}}$, a router induces the support labeling $\{I_q(x_i)\}_{i=1}^{n_{\mathrm{tr}}}$. Because the support is allowed to vary with x_i, it is not sufficient to union-bound over the $\binom{M}{q}$ possible static supports. Instead, Assumption 2 and (A.1) bound the number of distinct support labelings realizable on the entire observed sample. Consequently, the logarithmic contribution of hard support selection is of order $q\log(eM/q)$.

Conditional on any one realizable support labeling, the active masked softmax is smooth. Since $\tau \ge \tau_{\min} > 0$, its Jacobian with respect to the active logits is uniformly bounded by a constant multiple of $\tau_{\min}^{-1}$. Hence the active routing weights are Lipschitz functions of the router logits. Under the boundedness conditions in Assumption 2, the map

$$(h_0, \{\bar{h}_e : e \in I_q\}, \{\pi_e : e \in I_q\}) \longmapsto h_0 + \gamma \sum_{e\in I_q} \pi_e \bar{h}_e$$

is also uniformly Lipschitz. The admissible prediction head and score map are Lipschitz on the same bounded range. Standard contraction of covering numbers therefore propagates the entropy bound for the router logits through the masked softmax, routed representation, prediction head, and score map. Together with the above support-growth contribution, this gives, uniformly over empirical measures Q,

$$\log N\left(\epsilon, \mathcal{V}_\lambda, L_2(Q)\right) \le C\left\{q\log(eM/q) + V_r \log(A/\epsilon)\right\}, \qquad 0 < \epsilon \le 1,$$

where the bounded-complexity backbone, head, and transport contributions are included in the constant in Assumption 2. The important point is that the previous equation controls the actual input-dependent top-q class, with the support allowed to vary across inputs instead of being fixed globally.

We next pass from entropy to stochastic error. Let $\mathcal{G}_\lambda = \{\ell_V - \ell_{V_\lambda^\circ} : V \in \mathcal{V}_\lambda\}$ be the corresponding centered excess-loss class, where ℓ_V denotes the per-observation differentiable training loss. Symmetrization gives

$$\mathbb{E}\sup_{g\in\mathcal{G}_\lambda(r)} \left|(\mathbb{P}_{n_{\mathrm{tr}}} - P_0)g\right| \le 2\mathbb{E}\sup_{g\in\mathcal{G}_\lambda(r)} \left|\frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} \epsilon_i g(Z_i)\right|,$$

where ϵ_i are independent Rademacher variables and $\mathcal{G}_\lambda(r)$ denotes the class localized to the neighborhood specified in Assumption 4. Applying the entropy bound in the previous equation to the usual localized chaining integral yields, up to constants,

$$\mathbb{E}\sup_{g\in\mathcal{G}_\lambda(r)} \left|(\mathbb{P}_{n_{\mathrm{tr}}} - P_0)g\right| \lesssim r\sqrt{\frac{q\log(eM/q) + V_r\log(An_{\mathrm{tr}})}{n_{\mathrm{tr}}}} + \frac{q\log(eM/q) + V_r\log(An_{\mathrm{tr}})}{n_{\mathrm{tr}}}.$$

The same bound holds with probability at least $1-\delta$ after adding the usual $\log(1/\delta)$ term; see, for example, van der Vaart and Wellner (1996) and Wainwright (2019).

Let $\Xi_\lambda(\delta) = q\log(eM/q) + V_r\log(An_{\mathrm{tr}}) + \log(1/\delta)$. The Bernstein/local quadratic-margin condition in (A.3) relates the variance and curvature of the localized excess-loss class to the score distance. Applying the standard localized fixed-point argument to the above equation therefore gives

$$\mathcal{L}_\lambda(\tilde{V}_\lambda) - \mathcal{L}_\lambda(V_\lambda^\circ) \le C\frac{\Xi_\lambda(\delta)}{n_{\mathrm{tr}}}$$

with probability at least $1-\delta$. The quadratic part of (A.3) then implies

$$\|\tilde{V}_\lambda - V_\lambda^\circ\|_{L_2(P_0\times\mathrm{Unif}[C])} \le C\sqrt{\frac{\Xi_\lambda(\delta)}{n_{\mathrm{tr}}}}.$$

Thus the square-root term in (A.6)is the score-scale consequence of the fast excess-objective rate.

It remains to account for the fact that the implemented estimator differs from the ideal exact minimizer just analyzed. Assumption 4 defines $\varepsilon_{\mathrm{opt}}$ on the score scale, so the approximate optimizer satisfies

$$\|\hat{V}_\lambda - \tilde{V}_\lambda\|_{L_2(P_0\times\mathrm{Unif}[C])} \le C\varepsilon_{\mathrm{opt}}.$$

The remaining perturbation comes from the estimated source experts and transports. Lemma A.1 gives

$$\sum_{e\in\mathcal{E}^*} \omega_e \|\hat{T}_e\hat{h}_e - T_e^* h_e^*\|_{L_2(P_0)} \le \varepsilon_T,$$

and the bounded Lipschitz propagation through the routed representation, prediction head, and score map transfers this perturbation to the score scale

$$\|V_\lambda^{\text{est source}} - V_\lambda^{\text{population source}}\|_{L_2(P_0\times\mathrm{Unif}[C])} \le C\varepsilon_T.$$

Combining the above equations with the triangle inequality yields

$$\|\hat{V}_\lambda - V_\lambda^\circ\|_{L_2(P_0\times\mathrm{Unif}[C])} \le C\left[\sqrt{\frac{q_\lambda \log(eM/q_\lambda) + V_r\log(An_{\mathrm{tr}}) + \log(1/\delta)}{n_{\mathrm{tr}}}} + \varepsilon_T + \varepsilon_{\mathrm{opt}}\right].$$

This proves the lemma.

□

Lemma A.3 controls the estimation error for each fixed candidate. The implemented procedure, however, must also choose among the finitely many prespecified candidates using the independent validation sample while respecting the coverage and computation guards. The next lemma quantifies the additional price of this data-dependent selection step and shows that, under the feasible-oracle condition, validation does not change the target of comparison.

Lemma A.4. *Under Assumptions 5 and 6, with probability at least $1-\delta$, the candidate selected by the foldwise validation rule satisfies*

$$\mathcal{R}_\alpha(f_{\hat{\lambda}}^{\mathrm{pre}}, V_{\hat{\lambda}}^{\mathrm{pre}}) \le \min_{\lambda\in\Lambda} \mathcal{R}_\alpha(f_\lambda^{\mathrm{pre}}, V_\lambda^{\mathrm{pre}}) + C\sqrt{\frac{\log(|\Lambda|/\delta)}{n_{\mathrm{val}}}}. \tag{A.7}$$

Proof. Let the validation sample be partitioned in advance into a fixed number K_{val} of folds $\mathcal{I}_1,\dots,\mathcal{I}_{K_{\mathrm{val}}}$, with fold sizes of order $n_{\mathrm{val}}/K_{\mathrm{val}}$. Throughout the proof, condition on all fitted candidate models. Because candidate fitting uses no validation observations, the validation sample remains i.i.d. from P_0 under this conditioning. We fix $\lambda\in\Lambda$ and a validation fold k. Let $\hat{q}_{\lambda,-k}$ denote the ordinary empirical $(1-\alpha)$ score quantile computed from the validation observations outside $\mathcal{I}_k$, and let $q_\lambda = q_\alpha(f_\lambda^{\mathrm{pre}}, V_\lambda^{\mathrm{pre}})$ denote the corresponding population quantile. The observations used to construct $\hat{q}_{\lambda,-k}$ are independent of those in the held-out fold $\mathcal{I}_k$.

By the Dvoretzky-Kiefer-Wolfowitz inequality, followed by a union bound over the finite candidate grid and the fixed validation folds, with probability at least $1-\delta/3$,

$$\sup_{\lambda\in\Lambda}\max_{1\le k\le K_{\mathrm{val}}}\sup_t \left|\hat{F}_{\lambda,-k}(t) - F_\lambda(t)\right| \le C\sqrt{\frac{\log(K_{\mathrm{val}}|\Lambda|/\delta)}{n_{\mathrm{val}}}},$$

where F_λ is the population true-label score distribution of candidate λ. Since Assumption 5 gives a density bounded below by $c_- > 0$ in a fixed neighborhood of q_λ, the quantile inversion argument used in Lemma A.2 gives

$$\sup_{\lambda\in\Lambda}\max_{1\le k\le K_{\mathrm{val}}} |\hat{q}_{\lambda,-k} - q_\lambda| \le C\sqrt{\frac{\log(K_{\mathrm{val}}|\Lambda|/\delta)}{n_{\mathrm{val}}}}.$$

The set-boundary margin in Assumption 5 then implies uniformly in (λ, k) that replacing q_λ by $\hat{q}_{\lambda,-k}$ changes population mean set size by at most the same order

$$\left|E_{P_0}\left|\{c : V_\lambda^{\mathrm{pre}}(X,c) \le \hat{q}_{\lambda,-k}\}\right| - \mathcal{R}_\alpha(f_\lambda^{\mathrm{pre}}, V_\lambda^{\mathrm{pre}})\right| \le C\sqrt{\frac{\log(K_{\mathrm{val}}|\Lambda|/\delta)}{n_{\mathrm{val}}}}.$$

We next control evaluation on the held-out fold. Conditional on the observations outside $\mathcal{I}_k$, the threshold $\hat{q}_{\lambda,-k}$ is fixed and the observations in $\mathcal{I}_k$ are i.i.d. The prediction-set size of a held-out observation lies in $[0, C]$ and the coverage indicator lies in $[0, 1]$. Hoeffding's inequality and another union bound over (λ, k) therefore imply, with probability at least $1 - \delta/3$,

$$\sup_{\lambda,k}\left|\hat{R}_{\lambda,k} - E_{P_0}\left|\{c : V_\lambda^{\mathrm{pre}}(X,c) \le \hat{q}_{\lambda,-k}\}\right|\right| \le C\sqrt{\frac{\log(K_{\mathrm{val}}|\Lambda|/\delta)}{n_{\mathrm{val}}}},$$

and the corresponding foldwise empirical coverage differs from its conditional population value by the same order. The activated-expert computation is also uniformly bounded under the fixed candidate budget; hence its held-out empirical mean obeys an analogous concentration bound.

Because K_{val} is fixed, averaging the foldwise quantities does not change the rate. Combining previous equations yields a simultaneous event on which, for every $\lambda \in \Lambda$,

$$\left|\hat{R}_\lambda - \mathcal{R}_\alpha(f_\lambda^{\mathrm{pre}}, V_\lambda^{\mathrm{pre}})\right| \le \Delta_{\mathrm{val}}, \quad \Delta_{\mathrm{val}} = C\sqrt{\frac{\log(|\Lambda|/\delta)}{n_{\mathrm{val}}}},$$

after absorbing the fixed K_{val} into the constant. The empirical coverage and computation guards are uniformly within $O(\Delta_{\mathrm{val}})$ of their population counterparts on the same event.

Let $\lambda^\star \in \arg\min_{\lambda\in\Lambda} \mathcal{R}_\alpha(f_\lambda^{\mathrm{pre}}, V_\lambda^{\mathrm{pre}})$ denote a population oracle candidate. By Assumption 6, its population coverage and computation constraints satisfy the validation guards with fixed positive slack. For sufficiently large n_{val}, the uniform deviations above are smaller than this slack, so $\lambda^\star$ remains empirically feasible. Since $\hat{\lambda}$ minimizes foldwise empirical mean set size among the empirically feasible candidates, $\hat{R}_{\hat{\lambda}} \le \hat{R}_{\lambda^\star}$. Using the previous equation, we can get

$$\mathcal{R}_\alpha(f_{\hat{\lambda}}^{\mathrm{pre}}, V_{\hat{\lambda}}^{\mathrm{pre}}) \le \hat{R}_{\hat{\lambda}} + \Delta_{\mathrm{val}} \le \hat{R}_{\lambda^\star} + \Delta_{\mathrm{val}} \le \mathcal{R}_\alpha(f_{\lambda^\star}^{\mathrm{pre}}, V_{\lambda^\star}^{\mathrm{pre}}) + 2\Delta_{\mathrm{val}}.$$

Absorbing the factor two into C proves (A.7).

Note that if the strict population slack in Assumption 6 is removed, the same argument no longer guarantees that the unrestricted population oracle is empirically feasible. In that case the comparison is instead with the best candidate in the population-feasible set enlarged by the uniform coverage and computation deviations. The strict-slack condition is therefore precisely what permits the comparator in (A.7) to be the unrestricted finite-grid oracle.

□

Proof of Theorem 4.1. For each $\lambda \in \Lambda$, define the population conformal-risk oracle within the corresponding candidate class by

$$R_\lambda^\star = \inf_{(f,V)\in\mathcal{H}_\lambda} \mathcal{R}_\alpha(f, V) \text{ and } R^\star = \min_{\lambda\in\Lambda} R_\lambda^\star.$$

Let $(\tilde{f}_\lambda, \tilde{V}_\lambda)$ denote the candidate fitted on $\mathcal{D}_0^{\mathrm{tr}}$ and used in validation, and let $(\bar{f}_\lambda, \bar{V}_\lambda)$ denote the hypothetical refit of the same candidate specification on $\mathcal{D}_0^{\mathrm{tr}} \cup \mathcal{D}_0^{\mathrm{val}}$. Only the refit corresponding to the selected $\hat{\lambda}$ is actually computed, so $(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}}) = (\bar{f}_{\hat{\lambda}}, \bar{V}_{\hat{\lambda}})$. Introducing the hypothetical refits for all λ is only a proof device that permits the learning bound to hold uniformly before the data-dependent choice of $\hat{\lambda}$ is substituted.

We set $q_{\max} = \max_{\lambda\in\Lambda} q_\lambda$,

$$\Delta_{\mathrm{tr}} = \sqrt{\frac{q_{\max}\log(eM/q_{\max}) + V_r\log(An_{\mathrm{tr}}) + \log(|\Lambda|/\delta)}{n_{\mathrm{tr}}}} \text{ and } \Delta_{\mathrm{val}} = \sqrt{\frac{\log(|\Lambda|/\delta)}{n_{\mathrm{val}}}}.$$

Apply Lemma A.3 with failure probability proportional to $\delta/|\Lambda|$ simultaneously over the finite candidate grid. By the uniform-refit part of Assumption 4, the same argument applies simultaneously to the

hypothetical refits. A union bound therefore gives an event of probability at least $1-\delta/2$ on which the localized score bounds hold for every training-only fit and every hypothetical refit. We first connect the population target of the smoothed learner to the conformal-risk oracle. By the surrogate-to-risk condition in Assumption 4 and Lemma A.2,

$$R_\alpha(f_\lambda^\circ, V_\lambda^\circ) \le R_\alpha(f_\lambda^\star, V_\lambda^\star) + C\varepsilon_{\mathrm{smooth}} = R_\lambda^\star + C\varepsilon_{\mathrm{smooth}}, \quad \lambda \in \Lambda.$$

The localized learning bound of Lemma A.3, together with the source-transport and optimization perturbation bounds, gives uniformly over $\lambda \in \Lambda$,

$$R_\alpha(\tilde{f}_\lambda, \tilde{V}_\lambda) \le R_\alpha(f_\lambda^\circ, V_\lambda^\circ) + C\{\Delta_{\mathrm{tr}} + \varepsilon_T + \varepsilon_{\mathrm{opt}}\},$$

and the same inequality holds for the hypothetical refit $(\bar{f}_\lambda, \bar{V}_\lambda)$. Combining the preceding two displays yields

$$R_\alpha(\tilde{f}_\lambda, \tilde{V}_\lambda) \le R_\lambda^\star + B_{\mathrm{tr}} \text{ and } R_\alpha(\bar{f}_\lambda, \bar{V}_\lambda) \le R_\lambda^\star + B_{\mathrm{tr}},$$

where $B_{\mathrm{tr}} = C\{\Delta_{\mathrm{tr}} + \varepsilon_T + \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{smooth}}\}$.

Independently, Lemma A.4, with its failure probability allocated as $\delta/2$, gives

$$\mathcal{R}_\alpha(\tilde{f}_{\hat{\lambda}}, \tilde{V}_{\hat{\lambda}}) \le \min_{\lambda\in\Lambda} \mathcal{R}_\alpha(\tilde{f}_\lambda, \tilde{V}_\lambda) + C\Delta_{\mathrm{val}}.$$

Assumption 6 is used to ensure that the population oracle candidate remains admissible under the validation coverage and computation guards, so the minimum in the above equation is compared with the unrestricted finite-grid oracle, not merely with an empirically feasible subcollection.

We now combine the learning and selection bounds. Because $(\tilde{f}_{\hat{\lambda}}, \tilde{V}_{\hat{\lambda}}) \in \mathcal{H}_{\hat{\lambda}}$, $R_{\hat{\lambda}}^\star \le \mathcal{R}_\alpha(\tilde{f}_{\hat{\lambda}}, \tilde{V}_{\hat{\lambda}})$. Hence, using previous equations, we can get

$$\begin{aligned}\mathcal{R}_\alpha(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}}) &= \mathcal{R}_\alpha(\bar{f}_{\hat{\lambda}}, \bar{V}_{\hat{\lambda}}) \le R_{\hat{\lambda}}^\star + B_{\mathrm{tr}} \le \mathcal{R}_\alpha(\tilde{f}_{\hat{\lambda}}, \tilde{V}_{\hat{\lambda}}) + B_{\mathrm{tr}} \\ &\le \min_{\lambda\in\Lambda} \mathcal{R}_\alpha(\tilde{f}_\lambda, \tilde{V}_\lambda) + B_{\mathrm{tr}} + C\Delta_{\mathrm{val}} \le \min_{\lambda\in\Lambda} R_\lambda^\star + 2B_{\mathrm{tr}} + C\Delta_{\mathrm{val}}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&R_\alpha(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}}) - \inf_{\lambda\in\Lambda} \inf_{(f,V)\in\mathcal{H}_\lambda} R_\alpha(f, V) \\ &\le C\left[\sqrt{\frac{q_{\max}\log(eM/q_{\max}) + V_r\log(An_{\mathrm{tr}}) + \log(|\Lambda|/\delta)}{n_{\mathrm{tr}}}}\right. \\ &\left. + \sqrt{\frac{\log(|\Lambda|/\delta)}{n_{\mathrm{val}}}} + \varepsilon_T + \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{smooth}}\right].\end{aligned}$$

after absorbing numerical factors and the fixed allocation of δ into C. This proves the oracle inequality.

If the routed candidate classes are misspecified relative to the unrestricted population target oracle, let $\mathcal{A}(\Pi_q)$ denote the approximation error from the partial-sharing condition. By bounded Lipschitz propagation from logits to scores and Lemma A.2, the approximation condition gives $R^\star \le R_0^\star + C\mathcal{A}(\Pi_{q_{\hat{\lambda}}})$, where $R_0^\star$ is the unrestricted population oracle risk. Substituting this inequality above adds $\mathcal{A}(\Pi_{q_{\hat{\lambda}}})$ to the right-hand side, as stated. The theorem concerns the predictor-score pair before certification; no certification error is used in this argument.

□

The oracle inequality separates the statistical cost of learning an input-dependent sparse route from the cost of finite-grid validation selection and source/transport estimation. Optimization and smoothing errors remain explicit and are not absorbed into the constant. This decomposition also distinguishes the roles of target and source sample sizes: increasing n_{tr} primarily improves route estimation, whereas larger source samples act through the source/transport term ε_T. Theorem 4.1 concerns the pre-certification predictor. For completeness, we define

$$\varepsilon_{\mathrm{cert}} = [\mathcal{R}_\alpha(\hat{f}^{\mathrm{cert}}, \hat{V}^{\mathrm{cert}}) - \mathcal{R}_\alpha(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}})]_+.$$

The corresponding post-certification bound is obtained by adding $\varepsilon_{\mathrm{cert}}$ to the right-hand side of Theorem 4.1. Proposition 4.3 controls harmful retained expert-cell pairs for the bounded safety loss

but, without an additional certification-power condition, does not upper-bound $\varepsilon_{\mathrm{cert}}$. In particular, an under-supported but genuinely useful expert may be rejected, increasing set size despite the validity of the safety certificate. Thus the certification guarantee should not be interpreted as automatically pre-serving the pre-certification efficiency bound. The finite calibration sample contributes a separate efficiency term. Conditional on the frozen fitted score and under the i.i.d. part of Assumption 1, the Dvoretzky-Kiefer-Wolfowitz inequality gives

$$\sup_t \left|\hat{F}_{\mathrm{cal}}(t) - F(t)\right| = O_{\mathbb{P}}(n_{\mathrm{cal}}^{-1/2}).$$

The positive score density in Assumption 5 then yields

$$|\hat{q}_{1-\alpha} - q_{1-\alpha}| = O_{\mathbb{P}}(n_{\mathrm{cal}}^{-1/2}),$$

and the set-boundary margin implies

$$\mathbb{E}|\hat{C}_\alpha(X)| - \mathbb{E}|C_{\alpha,\hat{f},\hat{V}}(X)| = O_{\mathbb{P}}(n_{\mathrm{cal}}^{-1/2}).$$

This efficiency rate uses i.i.d. calibration data; the exact finite-sample coverage result proved below requires only calibration-test exchangeability.

We now turn to the separation result. The following construction isolates the structural limitation of whole-source transfer and shows why input-dependent routing can attain strictly smaller conformal risk. Consider binary classification with two target regions G_1 and G_2 having probabilities $p_1, p_2 \geq \rho$, where $\alpha < \rho \leq 1/2$, and let $Y = 1$ almost surely. We fix $0 < \epsilon < 1/2$. The target-only predictor assigns probability $1/2$ to each class. Source 1 contains an expert that is useful on G_1 and acts as a shortcut on G_2, whereas Source 2 has the reverse pattern. A globally accepted source contributes a fixed signed logit margin, so any nonnegative source-level weights (w_1, w_2) shared across target inputs induce margins proportional to $w_1 - w_2$ on G_1 and $w_2 - w_1$ on G_2. Thus, if $w_1 \neq w_2$, one region of probability at least ρ receives the shortcut sign while the other receives the useful sign; if $w_1 = w_2$, both regions reduce to the ambiguous target-only predictor. The construction can be realized with two experts per source, one matched and one shortcut, with shortcut magnitude chosen so that the resulting true-label probabilities are at most ϵ on the harmed region and at least $1 - \epsilon$ on the matched region, giving probability gap $\kappa = 1 - 2\epsilon$. The corresponding local oracle routes Source 1 on G_1 and Source 2 on G_2. For LAC this produces true-label score ϵ and competing-label score $1 - \epsilon$ throughout the target population, so its population conformal risk is $\mathcal{R}^*_\alpha = 1$. Theorem 4.2 shows that every whole-source rule retains a strictly positive risk gap because it must use the same source-level decision across both regions.

Proof of Theorem 4.2. We fix any possibly data-dependent whole-source rule $\hat{s}$ and condition on the realized training data. Under this conditioning, its global source weights (w_1, w_2) are fixed. The argument below is therefore deterministic conditional on the training sample.

First suppose $w_1 = w_2$. By construction, the useful and shortcut contributions cancel in both regions, so the resulting predictor assigns probability $1/2$ to each class on G_1 and G_2. For LAC, both candidate-label scores are therefore $1/2$. The population $(1 - \alpha)$ quantile is $1/2$, both labels are included everywhere, and hence $\mathcal{R}_\alpha(\hat{s}) = 2$.

Now suppose $w_1 \neq w_2$. Relabel the two regions, if necessary, so that G_{g} is the region receiving the favorable signed margin and G_{b} is the region receiving the shortcut margin. The construction is symmetric across the two regions; if the true-label probability on G_{g} is p_{g}, then that on G_{b} is

$$p_{\mathrm{b}} = 1 - p_{\mathrm{g}}, \quad p_{\mathrm{g}} \geq 1 - \epsilon, \quad p_{\mathrm{b}} \leq \epsilon.$$

For binary LAC, the true-label and competing-label scores are consequently

	true label	competing label
G_{g}	p_{b}	p_{g}
G_{b}	p_{g}	p_{b}.

Thus the true-label score takes the lower value p_{b} on G_{g} and the higher value p_{g} on G_{b}.

By construction, $P_0(G_{\mathrm{b}}) \geq \rho > \alpha$. Hence

$$P_0\{V(X, Y) < p_{\mathrm{g}}\} \leq 1 - P_0(G_{\mathrm{b}}) < 1 - \alpha,$$

whereas $P_0\{V(X, Y) \leq p_{\mathrm{g}}\} = 1$. It follows directly from the definition of the population conformal

quantile that $q_\alpha(\hat{s}) = p_{\mathrm{g}}$. Both entries in each row of the above table are at most p_{g}, so both labels belong to the conformal set on both regions. Therefore $\mathcal{R}_\alpha(\hat{s}) = 2$ also when $w_1 \neq w_2$.

The argument is conditional on an arbitrary realized training sample and uses no restriction on how the whole-source rule obtained (w_1, w_2) other than that the same source weights are used at every target input. Consequently, $\mathbb{E}_P\{\mathcal{R}_\alpha(\hat{s}) \mid \mathcal{D}_{\mathrm{train}}\} = 2$ for the constructed distribution P, and hence $\mathbb{E}_P\{\mathcal{R}_\alpha(\hat{s}) - \mathcal{R}^*_\alpha\} = 1$, because the regionwise local oracle has $\mathcal{R}^*_\alpha = 1$ by the preceding calculation. Since this conclusion holds for every whole-source rule, including data-dependent ones,

$$\inf_{\hat{s}} \sup_{P \in \mathcal{P}_{\mathrm{part}}} \mathbb{E}_P\{\mathcal{R}_\alpha(\hat{s}) - \mathcal{R}^*_\alpha\} \geq 1$$

for the deterministic construction.

The same conclusion is robust to a sufficiently small smoothing of the construction. Indeed, choose the smoothing so that the region masses remain bounded below by ρ, the useful-shortcut probability separation remains bounded away from zero by the prescribed κ, and the induced score perturbation is at most r in the metric of Lemma A.2. That lemma changes each population conformal risk by at most $C_R r$. Therefore the deterministic gap of one becomes at least $1 - 2C_R r$. Taking r sufficiently small gives

$$\inf_{\hat{s}} \sup_{P \in \mathcal{P}_{\mathrm{part}}} \mathbb{E}_P\{\mathcal{R}_\alpha(\hat{s}) - \mathcal{R}^*_\alpha\} \geq c_0(\rho, \kappa) > 0,$$

which proves the asserted strict separation for the regularized family.

Finally, the local routing class contains the regionwise oracle by construction; it chooses the matched expert on G_1 and the other matched expert on G_2. Hence its approximation error is zero. Applying Theorem 4.1 with fixed routing budget and the present finite expert bank gives

$$\begin{aligned}
&\mathcal{R}_\alpha(\hat{f}^{\mathrm{pre}}, \hat{V}^{\mathrm{pre}}) - \mathcal{R}^*_\alpha \\
&= O_{\mathbb{P}}\left(\sqrt{\frac{V_r + \log M + \log(|\Lambda|/\delta)}{n_{\mathrm{tr}}}} + \sqrt{\frac{\log(|\Lambda|/\delta)}{n_{\mathrm{val}}}} + \varepsilon_T + \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{smooth}} \right),
\end{aligned}$$

up to constants depending only on the fixed regularity parameters. Thus the local-router regret vanishes whenever the displayed estimation, selection, source/transport, optimization, and smoothing terms vanish, whereas every whole-source rule retains the strictly positive gap above.

□

Theorem 4.2 is a structural statement and does not imply that local routing is favored in every empirical data set. Its role is to show that the transfer unit itself can create irreducible regret; even an arbitrarily sophisticated data-dependent selector cannot repair a source-level decision when useful and shortcut computations coexist inside the same source and exchange roles across target regions. The local router removes this obstruction because its support is allowed to depend on X.

We finally prove the certification statement. Conditional on $\mathcal{F}_{\mathrm{pre}}$, the proposed route, partition, reference threshold, and deletion operation are fixed. For a supported expert-cell pair (e, j), the certification observations contributing to that pair are i.i.d. draws from the conditional law given $X \in G_j$ and $e \in I_q(X)$. Since $0 \leq \ell_{\mathrm{safe}} \leq B_0$, the deletion contrast lies in $[-B_0, B_0]$. The empirical-Bernstein inequality of Maurer and Pontil (2009), applied after rescaling this interval and then union-bounded over at most MJ pairs, yields exactly the upper confidence bound in (3.4), with the displayed constant chosen conservatively.

Proof of the certification part of Proposition 4.3. Condition on all quantities fitted before the certification sample is opened, including the predictor, pre-certification router, partition $\mathcal{G} = \{G_j\}_{j=1}^J$, reference threshold, and deletion operation. Under this conditioning these objects are fixed, while the certification observations remain i.i.d. from P_0 by Assumption 1.

We fix an expert-cell pair (e, j) and write $A_{ej} = \{X \in G_j,\ e \in I_q(X)\}$. For certification observations satisfying A_{ej}, the deletion contrasts

$$D_{ie} = \ell_{\mathrm{safe}}(X_i, Y_i; \pi_i^{+e}) - \ell_{\mathrm{safe}}(X_i, Y_i; \pi_i^{-e})$$

are i.i.d. draws from the conditional distribution of $D_e(X, Y)$ given A_{ej}. Their conditional mean is

$\Delta_{ej} = \mathbb{E}\{D_e(X,Y) \mid A_{ej}\}$, and, because $0 \le \ell_{\text{safe}} \le B_0$, $-B_0 \le D_{ie} \le B_0$. Conditional also on $N_{ej} = m \ge n_{\min}$, the one-sided empirical-Bernstein inequality applied to these m bounded observations gives

$$P\left\{\Delta_{ej} > \bar{D}_{ej} + \sqrt{\frac{2\hat{s}_{ej}^2 L_\delta}{N_{ej}}} + \frac{14 B_0 L_\delta}{3(N_{ej}-1)} \,\middle|\, N_{ej} = m\right\} \le \frac{\delta}{MJ},$$

with $L_\delta = \log(2MJ/\delta)$. The displayed upper bound is exactly U_{ej} in (3.4). Since it holds for every admissible value of N_{ej}, removing the conditioning on N_{ej} preserves the same bound. A union bound over at most MJ expert-cell pairs therefore yields an event $\mathcal{E}_{\text{cert}}$ satisfying $P(\mathcal{E}_{\text{cert}}) \ge 1-\delta$ on which $\Delta_{ej} \le U_{ej}$ simultaneously for every pair with $N_{ej} \ge n_{\min}$.

By construction, an expert-cell pair is retained only if $U_{ej} \le 0$. Hence, on $\mathcal{E}_{\text{cert}}$, every retained supported pair satisfies $\Delta_{ej} \le U_{ej} \le 0$. Consequently,

$$P\{\exists (e,j) : U_{ej} \le 0,\ \Delta_{ej} > 0\} \le \delta,$$

which proves the simultaneous familywise certification statement.

It remains to establish the finite-partition bound. Whenever the relevant conditional expectations exist, define $\Delta_e(x) = \mathbb{E}\{D_e(X,Y) \mid X = x,\ e \in I_q(X)\}$. For a supported pair (e,j), $\Delta_{ej} = \mathbb{E}\{\Delta_e(X) \mid X \in G_j,\ e \in I_q(X)\}$. Therefore, for any $x \in G_j$ with $e \in I_q(x)$,

$$\Delta_e(x) \le \Delta_{ej} + \text{osc}_j, \quad \text{osc}_j = \sup_{\substack{x,x' \in G_j \\ e \in \mathcal{E}}} |\Delta_e(x) - \Delta_e(x')|.$$

If (e,j) is retained, then $\Delta_{ej} \le 0$ on $\mathcal{E}_{\text{cert}}$, and hence $[\Delta_e(x)]_+ \le \text{osc}_j$. At every target input at most q experts are proposed, and therefore at most q retained experts can contribute. Thus, for $x \in G_j$,

$$\sum_{\substack{e \text{ retained} \\ e \in I_q(x)}} [\Delta_e(x)]_+ \le q\ \text{osc}_j\,.$$

Integrating over X and summing over the partition cells gives

$$\mathbb{E}_{P_0}\left[\sum_{\substack{e \text{ retained} \\ e \in I_q(X)}} [\Delta_e(X)]_+\right] = \sum_{j=1}^{J} \mathbb{E}_{P_0}\left[\mathbf{1}\{X \in G_j\} \sum_{\substack{e \text{ retained} \\ e \in I_q(X)}} [\Delta_e(X)]_+\right] \le q \sum_{j=1}^{J} P_0(G_j)\,\text{osc}_j,$$

which is the partition-bias bound stated in Proposition 4.3.

□

The certification statement is intentionally one-deletion-at-a-time. Each D_{ie} compares the frozen pre-certification route with the counterfactual obtained by deleting only expert e and renormalizing the remaining weights. If several experts are subsequently removed, their effects need not add because the renormalized route changes after each deletion. Proposition 4.3 therefore certifies retained expert-cell decisions relative to this specified intervention; it does not assume additivity, submodularity, or joint optimality of the retained subset. Pairs with insufficient certification support are rejected from certification, trading efficiency for conservative safety.

A.3 Calibration Firewall and Scope of the Guarantees

Certification completes all adaptive transfer decisions before the calibration sample is opened. The remaining argument therefore conditions on the fully fitted and certified procedure and uses calibration only to determine the terminal conformal rank.

Proof of the coverage part of Proposition 4.3. Condition on $\mathcal{F}_{\text{fit}}$. By Assumption 1, all fitted predictors, transports, routing and certification decisions, score parameters, and preprocessing operations are then fixed, while $S_i = \hat{V}^{\text{cert}}(X_i, Y_i),\ i = 1, \dots, n_{\text{cal}},\ S_* = \hat{V}^{\text{cert}}(X_*, Y_*)$ are exchangeable. Let $k_\alpha = \lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$. If $k_\alpha \le n_{\text{cal}}$, let $\hat{q} = S_{(k_\alpha)}$ be the k_αth order statistic of the n_{cal} calibration scores; if $k_\alpha = n_{\text{cal}} + 1$, set $\hat{q} = +\infty$.

First suppose the $n_{\mathrm{cal}}+1$ scores are distinct. By exchangeability, the rank R_* of S_* among $S_1,\dots,S_{n_{\mathrm{cal}}},S_*$ is uniform on $\{1,\dots,n_{\mathrm{cal}}+1\}$. Whenever $R_*\le k_\alpha$, fewer than k_α calibration scores are smaller than S_*, and hence $S_*\le S_{(k_\alpha)}=\hat{q}$. Therefore, we have

$$P\{S_*\le \hat{q}\mid \mathcal{F}_{\mathrm{fit}}\}\ge P\{R_*\le k_\alpha\mid \mathcal{F}_{\mathrm{fit}}\}=\frac{k_\alpha}{n_{\mathrm{cal}}+1}\ge 1-\alpha.$$

Since $\hat{C}_\alpha(x)=\{y:\hat{V}^{\mathrm{cert}}(x,y)\le\hat{q}\}$, the event $S_*\le\hat{q}$ is exactly $Y_*\in\hat{C}_\alpha(X_*)$. Consequently,

$$P\{S_*\le\hat{q}\mid\mathcal{F}_{\mathrm{fit}}\}\ge 1-\alpha \text{ and } P\{Y_*\in\hat{C}_\alpha(X_*)\mid\mathcal{F}_{\mathrm{fit}}\}\ge 1-\alpha.$$

If ties occur, attach independent continuous tie breakers to the $n_{\mathrm{cal}}+1$ scores. The preceding rank argument applies to the tie-broken scores. Removing the tie breakers and retaining labels whenever the original score satisfies "$\le$" can only enlarge the resulting prediction set, so the same lower coverage bound holds without a continuity assumption on the score distribution. If $k_\alpha=n_{\mathrm{cal}}+1$, then $\hat{q}=+\infty$ and coverage is one. This is the standard split-conformal exchangeable-rank argument (Vovk et al., 2005; Lei et al., 2018).

□

The coverage guarantee in Proposition 4.3 is assumption-light: it remains valid regardless of whether the learned transfer mechanism is predictively useful. Complete source failure, wholesale certificate rejection, or target-only fallback may affect efficiency, but none changes the exchangeable-rank calculation once the deployed score has been frozen. The conditioning on $\mathcal{F}_{\mathrm{fit}}$ permits arbitrary source training, architecture choice, target fitting, validation selection, the prescribed single refit, partition construction, and certification before calibration is opened. None of the partial-sharing, entropy, transport, optimization, quantile-density, or certification assumptions is required for this exact coverage statement; those conditions enter only the efficiency or safety results. Conversely, using calibration outcomes to alter the predictor, route, partition, certificate, score parameter, or checkpoint would violate the conditioning argument.

The finite-calibration efficiency rate and the exact coverage guarantee address different questions. The former uses i.i.d. calibration sampling together with local score regularity to quantify the perturbation of expected set size caused by replacing the population quantile with a finite-sample order statistic. The latter requires only conditional calibration-test exchangeability and continues to hold even when all source experts are uninformative and certification falls back completely to target-only prediction. The population partial-sharing decomposition is likewise not assumed to be identifiable at the individual-module level. If two experts span the same predictive direction, their contributions may be redistributed without changing the induced target predictor. The theoretical object is therefore the predictor-score pair and its conformal risk, without requiring a uniquely identified expert decomposition. Route-recovery F1 is reported only in the controlled simulation, where the generating expert support is fixed and separated by construction. Finally, discontinuities induced by hard top-q routing are handled statistically through the support-growth condition (A.1); no global derivative of the hard support map is assumed. Optimization uses the straight-through approximation described in Appendix B and retains an explicit optimization-error term. Router-support ties are resolved deterministically, expert-cell pairs with insufficient certification support are rejected, and if no proposed expert survives certification the deployed predictor reduces exactly to the target-only model. These conventions ensure that the deployed score is a fixed, $\mathcal{F}_{\mathrm{fit}}$-measurable function before calibration begins.

B Optimization and Algorithmic Details

This part supplies the differentiable calculations, straight-through top-q derivatives, block-coordinate updates, and computational accounting omitted from Section 3. Sparse top-k routing follows the mixture-of-experts tradition (Shazeer et al., 2017; Riquelme et al., 2021; Fedus et al., 2022; Zhou et al., 2022); RouteCP differs by routing transferable source computations and applying an independent target-side certificate before final conformal calibration.

B.1 Training Objective and Differential Calculus

This section gives the calculations used by the implementation but omitted from the main paper. The purpose is twofold: to make every gradient entering the alternating optimizer explicit and to separate

the differentiable training surrogate from the exact, nondifferentiable operations used for validation, certification, and final calibration. Throughout this section all sums over training observations are over $\mathcal{D}_0^{\mathrm{tr}}$ unless indicated otherwise.

The differentiable quantile surrogate uses the Huber-smoothed pinball loss. For $u \in (0,1)$ and smoothing bandwidth $h > 0$, define the Huber approximation to $|r|$ by

$$H_h(r) = \begin{cases} \frac{r^2}{2h}, & |r| \le h, \\ |r| - \frac{h}{2}, & |r| > h, \end{cases}$$

and the smoothed pinball loss by

$$\rho_{u,h}(r) = \frac{1}{2} H_h(r) + (u - \tfrac{1}{2})r.$$

This convention converges pointwise to the ordinary pinball loss $\rho_u(r) = r\{u - \mathbf{1}(r < 0)\}$ as $h \downarrow 0$. Its first derivative is

$$\psi_{u,h}(r) := \rho'_{u,h}(r) = \begin{cases} u - 1, & r < -h, \\ u - \frac{1}{2} - \frac{r}{2h}, & -h \le r < 0, \\ u - \frac{1}{2} + \frac{r}{2h}, & 0 \le r \le h, \\ u, & r > h, \end{cases}$$

where the equivalent compact expression $\psi_{u,h}(r) = u - \frac{1}{2} + \frac{1}{2}\,\mathrm{clip}(r/h, -1, 1)$ is used in code. Away from the two knots $\pm h$ the second derivative is

$$\psi'_{u,h}(r) = \frac{1}{2h} \mathbf{1}\{|r| < h\}.$$

The training quantile $\tilde{q}_u$ minimizes

$$J_q(t; \vartheta) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \rho_{u,h_q}\{v_i(\vartheta) - t\} + \frac{\xi_q}{2} t^2, \quad u = 1 - \alpha,$$

where ϑ denotes all model parameters entering the score. Since $t \mapsto J_q(t; \vartheta)$ is ξ_q-strongly convex, the minimizer is unique even when all scores lie outside the Huber window. The first-order condition is

$$\Psi(t, \vartheta) = -\frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \psi_{u,h_q}\{v_i(\vartheta) - t\} + \xi_q t = 0. \tag{B.1}$$

At points where no residual equals $\pm h_q$,

$$\partial_t \Psi(t, \vartheta) = \frac{1}{n_{\mathrm{tr}}} \sum_i \psi'_{u,h_q}\{v_i(\vartheta) - t\} + \xi_q \ge \xi_q > 0.$$

The implicit-function theorem therefore gives

$$\nabla_\vartheta \tilde{q}_u = -\{\partial_t \Psi(\tilde{q}_u, \vartheta)\}^{-1} \nabla_\vartheta \Psi(\tilde{q}_u, \vartheta) = \frac{\frac{1}{n_{\mathrm{tr}}} \sum_i \psi'_{u,h_q}(v_i - \tilde{q}_u) \nabla_\vartheta v_i}{\xi_q + \frac{1}{n_{\mathrm{tr}}} \sum_i \psi'_{u,h_q}(v_i - \tilde{q}_u)}. \tag{B.2}$$

Thus the derivative is a normalized average of score derivatives for observations whose residuals lie in the Huber window. The denominator is bounded below by ξ_q, which prevents the unstable derivative that can arise when the empirical score density near the desired quantile is nearly zero. At a Huber knot, any consistent subderivative may be used; the event has probability zero under a continuous score distribution and is immaterial for the theoretical statements.

The smooth set-size term then differentiates through that quantile. For LAC, we write $S_{ic}(\vartheta) = V_\eta(X_i, c; \vartheta)$ and define

$$\omega_{ic} = s\left\{\frac{\tilde{q}_u - S_{ic}}{h_{\mathrm{set}}}\right\}, \quad s(a) = \{1 + e^{-a}\}^{-1}.$$

Then $\hat{L}_{\mathrm{conf}} = n_{\mathrm{tr}}^{-1} \sum_{i,c} \omega_{ic}$. Since $s'(a) = s(a)\{1 - s(a)\}$,

$$\nabla_\vartheta \hat{L}_{\mathrm{conf}} = \frac{1}{n_{\mathrm{tr}} h_{\mathrm{set}}} \sum_{i,c} \omega_{ic}(1 - \omega_{ic})\{\nabla_\vartheta \tilde{q}_u - \nabla_\vartheta S_{ic}\}. \tag{B.3}$$

Equation (B.3), together with (B.2), is the exact gradient used by the differentiable quantile implementation. In a memory-constrained implementation, $\tilde{q}_u$ is treated as an implicit scalar node and its vector-Jacobian product is evaluated without materializing the full Jacobian.

For APS, order the class probabilities $p_{(1)}(x) \geq \cdots \geq p_{(C)}(x)$ and let $R_y(x)$ be the rank of label y. The deterministic APS score used in our experiments is

$$V_{\mathrm{APS}}(x, y) = \sum_{r=1}^{R_y(x)} p_{(r)}(x),$$

with a fixed deterministic tie rule. RAPS adds the regularization term

$$V_{\mathrm{RAPS}}(x, y) = V_{\mathrm{APS}}(x, y) + \lambda_{\mathrm{RAPS}}\{R_y(x) - k_{\mathrm{reg}}\}_+.$$

During gradient computation, the current probability ordering is treated as locally fixed, exactly as the top-q support is treated as locally fixed. This gives a piecewise differentiable objective. Final calibration and evaluation use the exact deterministic score and do not use this local-order approximation.

On a fixed top-q support, the masked-softmax derivatives are explicit. Let $A = I_q(x)$ be the support selected by the forward pass and write

$$\pi_e = \frac{\exp(r_e/\tau)}{\sum_{j \in A} \exp(r_j/\tau)}, \quad e \in A.$$

Conditional on A, the masked-softmax Jacobian is

$$\frac{\partial \pi_e}{\partial r_j} = \frac{1}{\tau}\pi_e\{\mathbf{1}(e = j) - \pi_j\}, \quad e, j \in A, \tag{B.4}$$

and is set to zero for $e \notin A$ or $j \notin A$ in the straight-through backward approximation. Hence, for any scalar downstream loss L,

$$\frac{\partial L}{\partial r_j} = \frac{1}{\tau}\pi_j \left\{ \frac{\partial L}{\partial \pi_j} - \sum_{e \in A} \pi_e \frac{\partial L}{\partial \pi_e} \right\}.$$

The logits satisfy $r_e = a^\top \sigma(\zeta_e)$ with $\zeta_e = A_0 h_0(x) + A_u u_e + c$. Therefore

$$\nabla_a r_e = \sigma(\zeta_e), \quad \nabla_{A_0} r_e = \{a \odot \sigma'(\zeta_e)\} h_0(x)^\top,$$
$$\nabla_{A_u} r_e = \{a \odot \sigma'(\zeta_e)\} u_e^\top, \quad \nabla_{u_e} r_e = A_u^\top \{a \odot \sigma'(\zeta_e)\}, \quad \nabla_c r_e = a \odot \sigma'(\zeta_e).$$

These expressions make clear that the expert embedding affects only the routing logit; it does not alter the source expert output itself.

The routed representation propagates gradients through the transfer amplitude, routing weights, and prediction head. For a fixed support, $z = h_0 + \gamma \sum_{e \in A} \pi_e \bar{h}_e$. Thus

$$\frac{\partial z}{\partial \gamma} = \bar{h}_\pi, \quad \frac{\partial z}{\partial \pi_e} = \gamma \bar{h}_e, \quad \frac{\partial z}{\partial \bar{h}_e} = \gamma \pi_e I_d.$$

If g_θ is a linear head, $g_\theta(z) = W_g z + b_g$, then $\nabla_z \ell = W_g^\top \nabla_g \ell$. For a nonlinear head, automatic differentiation applies to g_θ and the displayed identities remain the interface to the routed representation. To impose $\gamma \in [0, \gamma_{\max}]$, the implementation uses $\gamma = \gamma_{\max} s(\gamma_0)$, so $d\gamma/d\gamma_0 = \gamma_{\max} s(\gamma_0)\{1 - s(\gamma_0)\}$.

Affine transport and layer normalization are differentiated directly. Let $w_e = W_e h_e(x) + b_e$ and $\bar{h}_e = \mathrm{LN}(w_e)$. We write $\mu = d^{-1}\mathbf{1}^\top w_e$, $\tilde{w} = w_e - \mu\mathbf{1}$, and $s_e = (d^{-1}\|\tilde{w}\|_2^2 + \epsilon_{\mathrm{LN}})^{1/2}$. With unit gain and zero offset, $\mathrm{LN}(w_e) = \tilde{w}/s_e$. Its Jacobian is

$$J_{\mathrm{LN}}(w_e) = \frac{1}{s_e}\left(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top\right) - \frac{1}{d s_e^3}\tilde{w}\tilde{w}^\top. \tag{B.5}$$

Consequently, for upstream gradient $g_e = \nabla_{\bar{h}_e} L$,

$$\nabla_{W_e} L = J_{\mathrm{LN}}(w_e)^\top g_e \, h_e(x)^\top \text{ and } \nabla_{b_e} L = J_{\mathrm{LN}}(w_e)^\top g_e.$$

If affine gain and offset are used in layer normalization, their ordinary coordinatewise gradients are added. The small constant ϵ_{LN} is fixed before tuning.

The disagreement factor is differentiated from its weighted-variance representation. With $m = \bar{h}_\pi = \sum_e \pi_e \bar{h}_e$, the weighted dispersion can be written

$$d^2 = \sum_e \pi_e \|\bar{h}_e\|_2^2 - \|m\|_2^2.$$

On a fixed support,

$$\frac{\partial d^2}{\partial \pi_j} = \|\bar{h}_j\|_2^2 - 2m^\top \bar{h}_j \text{ and } \nabla_{\bar{h}_j} d^2 = 2\pi_j(\bar{h}_j - m).$$

For numerical stability we use $d = (d^2 + \epsilon_d)^{1/2}$ during training, so $\nabla d = (2d)^{-1}\nabla d^2$. The reported test-time score uses the same fixed ϵ_d. For LAC, $V_\eta(x, y) = \{1 - p_y(x)\}\{1 + \eta d(x)\}$, so

$$\nabla_\vartheta V_\eta = -(1 + \eta d)\nabla_\vartheta p_y + (1 - p_y)\{d\nabla_\vartheta \eta + \eta \nabla_\vartheta d\}.$$

When η is represented by $\eta = \mathrm{softplus}(\eta_0)$, its derivative is $s(\eta_0)$.

The remaining regularizers have standard moment and simplex derivatives. For class c and expert e, define $\hat{\mu}_{e,c}^{(s)} = n_{s,c}^{-1} \sum_{i:Y_{si}=c} \bar{h}_e(X_{si})$ and the corresponding target moment $\hat{\mu}_c^{(0)}$. The mean component of the alignment loss is $\|\hat{\mu}_{e,c}^{(s)} - \hat{\mu}_c^{(0)}\|_2^2$, whose derivative with respect to a source aligned representation is $\frac{2}{n_{s,c}}\{\hat{\mu}_{e,c}^{(s)} - \hat{\mu}_c^{(0)}\}$. For the covariance term $\|\hat{\Sigma}_e^{(s)} - \hat{\Sigma}^{(0)}\|_F^2$, let $D_e = \hat{\Sigma}_e^{(s)} - \hat{\Sigma}^{(0)}$. Differentiating the unbiased covariance gives, for a centered source observation $\tilde{h}_i = \bar{h}_e(X_{si}) - \hat{\mu}_e^{(s)}$,

$$\nabla_{\bar{h}_e(X_{si})} \|D_e\|_F^2 = \frac{4}{n_s - 1} D_e \tilde{h}_i,$$

with the analogous negative contribution for target observations entering $\hat{\Sigma}^{(0)}$.

For entropy $H(\pi) = -\sum_e \pi_e \log(\pi_e + \epsilon_H)$,

$$\frac{\partial H}{\partial \pi_e} = -(\log(\pi_e + \epsilon_H) + \pi_e/(\pi_e + \epsilon_H)).$$

The balancing term is $M \sum_e (\bar{\pi}_e - M^{-1})^2$ with $\bar{\pi}_e = n_{\mathrm{tr}}^{-1} \sum_i \pi_e(X_i)$, giving

$$\frac{\partial \hat{L}_{\mathrm{bal}}}{\partial \pi_e(X_i)} = \frac{2M}{n_{\mathrm{tr}}}(\bar{\pi}_e - M^{-1}).$$

All these derivatives are propagated through (B.4).

B.2 Optimization, Complexity, and Implementation Checks

One epoch consists of the following blocks. First update the target head and, when allowed by the candidate, the target backbone using cross-entropy plus the conformal surrogate. Second update all active affine transports using prediction and alignment terms. Third update router parameters using prediction, conformal, entropy, and balancing terms. Fourth update γ and η. The top-q support is recomputed in every forward pass. Temperature follows the deterministic schedule $\tau_t = \max\{\tau_{\min}, \tau_0 \rho_\tau^t\}$, $\tau_0 = 1$, $\tau_{\min} = 0.1$, with ρ_τ chosen so that the floor is reached during the final quarter of training. A candidate is stopped at its prespecified epoch budget or earlier if the moving average of the training objective changes by less than ϵ_L and the support-change frequency is below ϵ_A for K_{stable} consecutive evaluations. The stopping rule is fixed before validation is inspected.

The data-flow display in Figure B.1 makes explicit why the learning, safety, and validity arguments are separated in Appendix A. Training and validation may change the fitted predictor, certification may only delete or renormalize proposed routes after the final refit, and calibration changes only the terminal order statistic. This ordering is also the implementation invariant checked in Table B.1.

The deployed computation follows from support-before-evaluation routing. Let F_0 be target-backbone FLOPs, F_r router-logit FLOPs, F_e source-expert FLOPs, F_{T_e} transport FLOPs, and F_g head FLOPs. If the selected support is $I_q(x)$ and the certified surviving support is $A_{\mathrm{cert}}(x) \subseteq I_q(x)$, deployment cost under early deletion is

$$F(x) = F_0 + F_r + \sum_{e \in A_{\mathrm{cert}}(x)} (F_e + F_{T_e}) + F_g.$$

When the certificate lookup is performed before expensive expert evaluation, rejected experts incur only router and lookup cost. If an implementation must evaluate all proposed experts before applying the certificate, replace $A_{\mathrm{cert}}(x)$ by $I_q(x)$ in the FLOP count.

Stage	Data newly opened	Permitted statistical use	Object frozen before the next stage
1. Learning	$\{\mathcal{D}_s\}_{s=1}^{S}$, $\mathcal{D}_0^{\mathrm{tr}}$	Fit the candidate transports, router, predictor, transfer amplitude, disagreement parameter, and score parameters for every $\lambda \in \Lambda$. Calibration and certification observations are inaccessible.	Fitted candidate family $\{\hat{m}_\lambda : \lambda \in \Lambda\}$.
2. Selection/refit	$\mathcal{D}_0^{\mathrm{val}}$	Evaluate candidates by fixed-fold ordinary conformal sets; select $\hat{\lambda}$ subject to the prespecified coverage and FLOP guards; then perform the single refit on $\mathcal{D}_0^{\mathrm{tr}} \cup \mathcal{D}_0^{\mathrm{val}}$.	Pre-certification predictor, selected score specification, partition rule $\mathcal{G}$, and reference threshold $\bar{q}$.
3. Certification	$\mathcal{D}_0^{\mathrm{cert}}$	With network parameters fixed, compute deletion contrasts D_{ie} and bounds U_{ej}; retain certified expert-cell pairs, delete unsupported pairs, renormalize surviving weights, and use target-only fallback when necessary.	Certified route, deployed predictor and score, certificate table, and $\hat{\eta}$.
Calibration firewall: all fitted, selected, and certified objects above are frozen before calibration is opened.			
4. Calibration	$\mathcal{D}_0^{\mathrm{cal}}$	Evaluate only the already-frozen deployed score on calibration observations and compute the split-conformal order statistic $k_\alpha = \lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil$.	Final threshold $\hat{q}_{1-\alpha}$ and prediction set $\hat{C}_\alpha(x)$.

Figure B.1: **Data-use ledger for RouteCP.** Each target split has one statistical role; training fits candidate models, validation selects and permits the single refit, certification modifies only the already-refitted route, and calibration determines only the terminal conformal rank. The firewall therefore separates adaptive transfer from the exchangeable rank calculation without suggesting any reverse or cross-stage information flow.

Algorithm B.1 Gradient-level training for one candidate

Require: Candidate λ, training data, source expert banks, epoch budget T.
Initialize target-only predictor, affine transports, router logits, γ, and η.
for $t = 1, \dots, T$ **do**
  Set $\tau_t = \max\{\tau_{\min}, \tau_0 \rho_\tau^t\}$.
  for minibatch B **do**
    Compute $h_0(x)$ and all lightweight router logits $r_e(x)$ for $x \in B$.
    Determine deterministic top-q support $I_q(x)$ and evaluate only selected source experts.
    Form masked-softmax weights, aligned representations, $z(x)$, probabilities, disagreement, and scores.
    Solve the one-dimensional strongly convex problem for $\tilde{q}_{1-\alpha}$.
    Evaluate $\hat{L}_{\mathrm{ce}}, \hat{L}_{\mathrm{conf}}, \hat{L}_{\mathrm{align}}, \hat{L}_{\mathrm{ent}}, \hat{L}_{\mathrm{bal}}$.
    Backpropagate through (B.2), treating the top-q support as locally fixed.
    Apply the prespecified optimizer step to the current parameter block.
  end for
  Record objective, support-change frequency, activated FLOPs, and fallback-free target-only comparator.
  if the prespecified stability rule is satisfied **then**
    break.
  end if
end for
return Fitted candidate and stopping epoch.

The router computes M lightweight logits. With embedding dimension d_u and hidden router width d_r, its leading cost is $O(Md_r)$ after the target representation has been computed. This is typically much

smaller than evaluating M source experts. Memory at inference stores expert embeddings $O(Md_u)$, router parameters, the active expert weights, and the certificate table. The latter has at most MJ scalar upper bounds plus support counts and is negligible relative to model weights. Training cost is higher because source and target moment alignment may require source minibatches and because all candidates in Λ are fitted. If T_λ is the epoch count for candidate λ, the rough training cost is

$$\sum_{\lambda \in \Lambda} T_\lambda n_{\mathrm{tr}} \left\{ F_0 + F_r + q_\lambda \bar{F}_e + F_g + F_{\mathrm{align}} \right\},$$

where $\bar{F}_e$ is average selected-expert cost. Candidate fits are embarrassingly parallel. The certification pass costs one forward evaluation for the proposed route plus up to q deletion reevaluations per certification observation; cached expert outputs make these deletion computations inexpensive because the network weights are frozen. Certification itself adds only a table lookup and post-routing deletion step at deployment. At deployment, the input is assigned to one partition cell using predicted class, confidence bin, and disagreement bin. The certificate is a table lookup for each proposed expert. Hence certification adds $O(q)$ scalar operations after cell assignment. It is not a neural critic evaluated at test time. The FLOP reduction translates into latency because the reported expert units are independently callable after the router selects them. FLOP savings translate to latency only if selected experts can be evaluated without materializing the full source bank. This is why the main router makes its support decision from $h_0(x)$ and expert embeddings before evaluating source experts. For architectures in which an "expert" is an intermediate block requiring execution of all preceding blocks, implementation should cache shared prefixes or define experts at independently callable adapter/LoRA/residual branches. The reported expert banks satisfy this callable-unit requirement.

Table B.1 records the direct correspondence between theoretical objects and implementation checks.

Table B.1: Correspondence between Mathematical Objects and Implementation Quantities.

Object	Definition in theory	Implementation check
$I_q(x)$	deterministic top-q support of router logits	stable sort with fixed expert-index tie breaker
$\pi_e(x)$	masked softmax on $I_q(x)$	zero outside support; sums to one within tolerance
$\tilde{q}_{1-\alpha}$	strongly convex smoothed pinball minimizer	scalar root/convex solve; gradient checked by finite differences
$\hat{L}_{\mathrm{conf}}$	smooth set-size surrogate	compared with exact set size as bandwidth decreases
$\varepsilon_{\mathrm{opt}}$	empirical optimization gap reported	report final objective and restart variability; never set to zero by assertion
$\mathcal{D}_0^{\mathrm{val}}$	finite-grid candidate selection data	no gradient updates from validation folds
$\mathcal{D}_0^{\mathrm{cert}}$	independent safety sample	loaded only after final refit; no parameter update afterward
U_{ej}	one-sided empirical-Bernstein upper bound	exact constants and $N_{ej}-1$ denominator unit tested
$\mathcal{D}_0^{\mathrm{cal}}$	final conformal rank sample	checksum that no training/tuning code accesses its labels
$\hat{C}_\alpha$	exact split-conformal set	deterministic $\leq$ inclusion and $+\infty$ edge case

Numerical gradient checks compare automatic derivatives with centered finite differences on small synthetic batches away from top-q and probability-order ties. For the implicit quantile derivative, a scalar model parameter is perturbed, the smoothed quantile is resolved at each perturbation, and the resulting finite-difference slope is compared with (B.2); relative errors above the prespecified tolerance fail the test suite. Separate unit tests enforce the data firewall. Training, validation, certification, and calibration are represented by distinct data-loader objects; calibration labels are monkey-patched while verifying that fitted weights, selected hyperparameters, partition boundaries, and certificate tables remain unchanged, so only the terminal conformal threshold can move. Certification labels may alter the certificate table but are verified not to update network parameters. These checks operationalize the conditioning structure used in Appendix A. We also rerun the complete reproducibility pipeline from a clean initialization and require the resulting numerical summaries, figures, and certificate decisions to agree with the archived outputs up to prespecified numerical tolerances. Each check is logged with seeds and tolerances. The final implementation additionally verifies that every reported configuration reproduces the data-access order in Algorithm 1 and that every table and figure is regenerated directly from stored seed-level outputs, with no manual transcription.

C Monte Carlo Design and Additional Simulations

This part gives the complete data-generating mechanism and additional simulation evidence supporting the numerical study in Section 5.1. We first specify the target regions, region-specific transferable experts, shortcut components, source transformations, and sample-splitting protocol used in the controlled partial-sharing design. We then report supplementary experiments that isolate routing-budget effects, source-sample scaling, certificate resolution, component ablations, and robustness to unsupported or misspecified transfer. All tuning grids, evaluation metrics, replication counts, and population-oracle quantities are fixed independently of the reported test outcomes, so the additional results diagnose the mechanisms studied in the theory and do not introduce post hoc simulation variants.

C.1 Complete Monte Carlo Design

The main simulation is a five-class problem with input dimension $p = 20$. We draw $X \sim \mathcal{N}(0, I_{20})$ and set

$$u = (1, 1, 1, 1, 0, \ldots, 0)^\top/2 \text{ and } Z = u^\top X \sim \mathcal{N}(0, 1).$$

Let $b_0 = -\infty$, $b_1 = \Phi^{-1}(.25)$, $b_2 = 0$, $b_3 = \Phi^{-1}(.75)$, and $b_4 = \infty$, and define $G_r = \{x : b_{r-1} < u^\top x \le b_r\}$ for $r \in [4]$. Thus $P_0(G_r) = 1/4$. There are $S = 4$ sources, four experts per source, and $M = 16$ source experts. Index the experts by $e = 4(s-1) + k$ for $s, k \in [4]$. The regionwise oracle supports are

$$A_1^* = \{1, 6\}, \quad A_2^* = \{3, 8\}, \quad A_3^* = \{10, 13\}, \quad A_4^* = \{12, 16\}.$$

Hence each region uses two experts from different sources and no complete source is uniformly optimal. In region G_r, the two experts in A_r^* are useful, six experts are neutral, and the remaining eight are shortcuts. To remove any ambiguity, the neutral sets are fixed as

$$N_1 = \{2, 3, 7, 10, 13, 16\}, \quad N_2 = \{1, 4, 6, 9, 12, 14\},$$
$$N_3 = \{2, 5, 8, 11, 14, 16\}, \quad N_4 = \{3, 5, 7, 9, 10, 13\}.$$

and $H_r = [16] \setminus (A_r^* \cup N_r)$ is the shortcut set. These sets are part of the DGP and are never supplied to the fitted router. Table C.1 records the complete role map, which makes the structural difficulty explicit. No source is uniformly useful across regions, while every region contains both reusable and misleading source computations. The default $q = 2$ matches the oracle support size, so routing performance reflects support discovery without the confounding effect of an artificially oversized budget.

Table C.1: **Complete Monte Carlo specification.** Panel A gives the fixed useful (U), neutral (N), and shortcut (H) roles for all 16 experts in each target region. Panel B records the default constants used throughout the simulation unless explicitly varied.

Panel A: expert roles by target region

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
G_1	U	N	N	H	H	U	N	H	H	N	H	H	N	H	H	N
G_2	N	H	U	N	H	N	H	U	N	H	H	N	H	N	H	H
G_3	H	N	H	H	N	H	H	N	H	U	N	H	U	N	H	N
G_4	H	H	N	H	N	H	N	H	N	N	H	U	N	H	H	U

Panel B: default design constants

Classes / input dimension	$C = 5, p = 20$	Sources / experts	$S = 4, K_s = 4, M = 16$
Representation dimension	$d_e = d = 32$	Routing budget	$q = 2$
Shortcut / feature noise	$\zeta = .8, \sigma_h = .15$	Certificate	$J = 8, n_{\min} = 20, \delta = .05$
Source sample	$n_s = 4000$ per source	Target train/val/cert/cal	800/400/400/800
Test size / coverage	$n_{\text{test}} = 5000, 1 - \alpha = .90$	Replications	500 performance; 1000 certificate

The remaining DGP is fixed deterministically up to the stated random draws. Let $v_c = e_c - 5^{-1} \sum_{\ell=1}^{5} e_\ell \in \mathbb{R}^{32}$ for $c \in [5]$, where e_c is the cth Euclidean basis vector, and define the centered private logit

$$f^*_{\text{private}}(x) = 0.35 \left[(x_5, x_6, x_7, x_8, x_9)^\top - 5^{-1} \Big\{ \sum_{j=5}^{9} x_j \Big\} \mathbf{1}_5 \right].$$

Region G_r has designated transferable class $c_r = r$ for $r \in [4]$. For the two useful experts in A_r^*, we set

$$f_e^*(x) = 0.85\{1 + 0.15 \tanh(x_{10+(e \bmod 6)})\}\, \tilde{v}_{c_r},$$

where $\tilde{v}_c$ is the first five coordinates of v_c. The target logit is $f_0^*(x) = f_{\text{private}}^*(x) + \sum_{e \in A_r^*} f_e^*(x)$ for $x \in G_r$, followed by centering across classes, and $Y \mid X = x \sim \text{Categorical}\{\text{softmax}(f_0^*(x))\}$. Neutral experts have zero region-specific transfer component. A shortcut expert $e \in H_r$ instead carries magnitude $\zeta = .8$ in the wrong class $\tilde{c}_{er} = 1 + \{c_r + e\} \bmod 5$, so it is predictive in its source construction but points away from the target-region class after transfer.

For the representation experiment, define

$$\psi(x) = (x_1, \ldots, x_{20}, \tanh x_1, \ldots, \tanh x_8, x_1x_2, x_3x_4, x_5x_6, x_7x_8)^\top \in \mathbb{R}^{32}.$$

The population aligned representation of expert e equals $\psi(x)$ plus its useful, neutral, or shortcut class-direction component embedded in the first five coordinates. The observed source representation is

$$h_e^{(s)}(x) = R_e D_e h_e^*(x) + \sigma_h \varepsilon_e, \quad \varepsilon_e \sim \mathcal{N}(0, I_{32}), \tag{C.1}$$

where $\sigma_h = .15$. For reproducibility, R_e is obtained by QR-orthogonalizing a 32×32 standard-normal matrix generated with PCG64 seed $202701 + e$, with the diagonal of the R factor made positive; D_e is diagonal with entries drawn independently from $\text{Unif}(.7, 1.3)$ using seed $202801 + e$. Fresh ε_e is drawn for each observation using the replication seed. The affine transport must therefore undo an expert-specific rotation and scale distortion in the presence of observation noise. Monte Carlo replication b uses master seed $271828 + b$, from which source, target, and optimization streams are spawned independently. These definitions make the simulation executable without relying on unstated coefficient matrices or hand-tuned region effects.

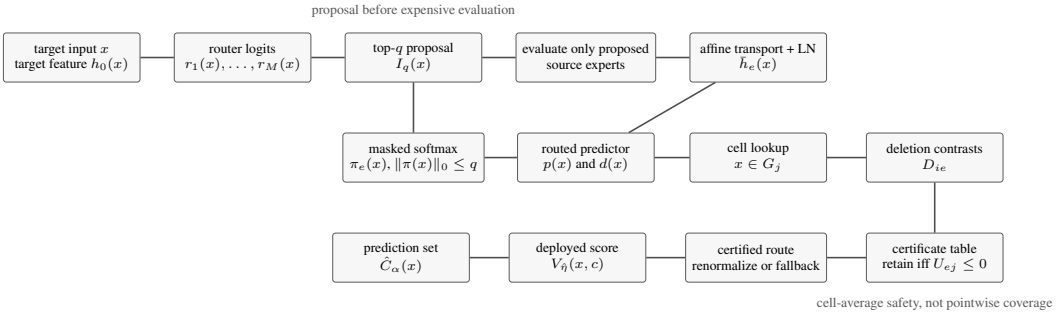


Figure C.1: **From a target input to a certified conformal set.** The router proposes a sparse support before source experts are evaluated, so sparsity corresponds to actual computation. Certification then acts on the already fitted route through the finite cell table and can delete unsupported experts without changing network parameters. Only the resulting certified predictor and score are passed to the final conformal rank calculation.

Figure C.1 links the estimator to the quantities reported in the simulation tables. Route F1 evaluates the top-q proposal against the region-specific oracle support, fallback records the event that certification removes every proposed expert, and harmful activation records retention of a adverse expert-cell contribution. Thus, the numerical diagnostics correspond to distinct stages of the deployed estimator and capture different aspects of transfer quality.

C.2 Sensitivity and Calibration Diagnostics

We next examine whether the observed gains persist under perturbations of the main routing and certification choices and whether the final conformal layer behaves as intended across nominal coverage levels. These diagnostics are designed to separate sensitivity of predictive efficiency from sensitivity of the validity mechanism, since the latter is protected by the untouched calibration split and does not depend on any particular routing specification.

Table C.2 shows that the default specification is not driven by a single fragile tuning choice. Increasing q yields only small additional reductions in set size while raising inference cost; moderate disagreement adjustment improves both set size and NT relative to $\eta = 0$; and larger $n_{\min}$ makes certification more conservative through higher fallback. Across nominal levels, empirical coverage tracks the target level closely. MCCD improves under RouteCP, but it remains a diagnostic and does not constitute a conditional-coverage guarantee. These results indicate that the main efficiency conclusions are stable over practically relevant tuning ranges, while the calibration firewall continues to determine marginal coverage independently of those tuning choices.

Table C.2: **Sensitivity and calibration diagnostics.** Panels A-C vary routing, disagreement, and certificate support; Panel D verifies marginal coverage across nominal levels; Panel E reports classwise coverage heterogeneity as a diagnostic only.

Panel A: routing budget					
q	Coverage	Size	NT	Relative FLOPs	–
1	.904	1.69	.021	1.24	–
2	.903	1.61	.014	1.44	–
4	.902	1.59	.015	1.73	–
8	.902	1.58	.018	2.31	–
Panel B: disagreement parameter					
η	Coverage	Size	MCCD	NT	–
0	.902	1.65	.051	.020	–
.25	.903	1.63	.048	.017	–
.50	.903	1.61	.045	.014	–
1.00	.905	1.66	.043	.013	–
Panel C: certificate minimum support					
$n_{\min}$	FA	Size	FR	Accepted pairs	–
10	.026	1.58	.092	41	–
20	.021	1.61	.147	34	–
30	.017	1.66	.205	28	–
40	.015	1.72	.289	21	–
Panel D: nominal coverage					
Method	80%	90%	95%	–	–
Target only	.804	.903	.953	–	–
Parameter-matched router	.801	.901	.951	–	–
RouteCP	.803	.903	.952	–	–
Panel E: classwise coverage diagnostic					
Benchmark	TO MCCD	RouteCP MCCD	Difference	–	–
PACS	.052	.037	-.015	–	–
OfficeHome	.073	.052	-.021	–	–
Camelyon17-WILDS	.081	.061	-.020	–	–
CLIP-PACS	.045	.033	-.012	–	–
CLIP-OfficeHome	.057	.041	-.016	–	–

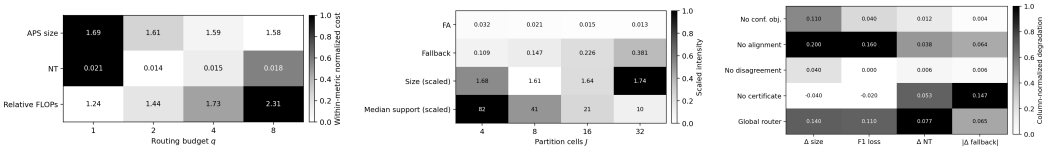
Table C.3 gives three complementary checks. Removing alignment produces the largest loss in route recovery, whereas removing certification gives the smallest sets but markedly increases negative transfer. As n_{tr} grows, local-routing excess risk continues to decline while the whole-source gap remains substantial, matching the separation result. The same ordering persists across independent replication batches. Certificate resolution exhibits the predicted bias-variance tradeoff; very fine partitions reduce false acceptance but increase fallback and eventually enlarge sets because many expert-cell pairs have limited support. This stability argues against conclusions driven by chance.

Table C.3: **Extended simulation results.** Panel A isolates the contribution of each component, Panel B traces the structural-separation pattern with target sample size, and Panel C studies the localization-support tradeoff of the finite partition.

Panel A: component ablations					
Variant	Coverage	Size	Route F1	NT	FR
Full RouteCP	.903	1.61	.83	.014	.147
No conformal objective	.902	1.72	.79	.026	.151
No alignment	.901	1.81	.67	.052	.211
No disagreement factor	.902	1.65	.83	.020	.141
No certificate	.901	1.57	.85	.067	0
Global router	.902	1.75	.72	.091	.082
Panel B: structural separation, excess conformal risk					
n_{tr}	Target only	Whole-source	Global gates	RouteCP	–
200	.287(.018)	.215(.016)	.146(.013)	.101(.011)	–
400	.253(.015)	.193(.014)	.111(.010)	.064(.008)	–
800	.234(.012)	.179(.011)	.083(.008)	.036(.006)	–
1600	.220(.010)	.171(.010)	.064(.007)	.021(.004)	–
3200	.212(.009)	.168(.009)	.054(.006)	.012(.003)	–
Panel C: finite-partition certificate sensitivity					
Cells J	Median N_{ej}	FA	NT	Size	FR
4	82	.032	.029	1.68	.109
8	41	.021	.014	1.61	.147
16	21	.015	.010	1.64	.226
32	10	.013	.009	1.74	.381

C.3 Extended Simulation Results

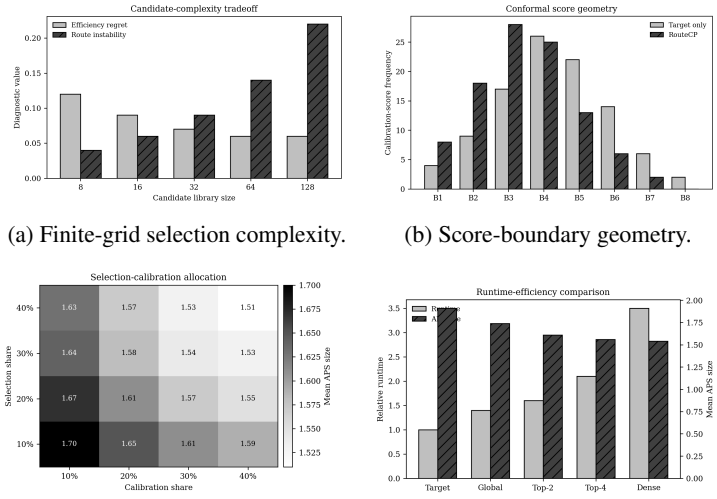
Figure C.2 condenses three numerical checks that are difficult to read from tables alone. The left panel shows why $q = 2$ is a meaningful default instead of the result of maximizing capacity; moving from $q = 2$ to $q = 8$ changes mean APS size from 1.61 to only 1.58 while relative FLOPs increase from 1.44 to 2.31. The middle panel visualizes the localization-support tradeoff behind Proposition 4.3; $J = 8$ is the point at which the reported set size is smallest before sparse cells drive fallback upward. The right panel separates the functions of the components: omitting alignment primarily damages support recovery, while omitting certification preserves route F1 but raises NT from $.014$ to $.067$.



(a) Routing-budget operating map. (b) Certificate resolution. (c) Ablation signature.

Figure C.2: **Operating mechanisms in controlled partial transfer.** Left: increasing q produces diminishing set-size gains while computation rises. Middle: finer certificate partitions lower false acceptance but reduce support and increase fallback. Right: normalized degradation relative to full RouteCP shows that alignment is most important for route recovery, whereas removing certification most strongly increases negative transfer. Numeric annotations are the underlying reported values.

Figure C.3 isolates the statistical mechanisms used in the proofs: finite-grid selection, score-boundary sensitivity, calibration-sample scaling, and the compute-efficiency frontier. Together with Tables C.2-C.3, they show that the reported gains persist across routing, certification, and sample-size perturbations and are not driven by a single default configuration.



(a) Finite-grid selection complexity. (b) Score-boundary geometry.

(c) Calibration-budget sensitivity. (d) Compute-efficiency frontier.

Figure C.3: **Statistical and computational scaling diagnostics.** (a) The finite-grid penalty grows only logarithmically with the number of prespecified candidates. (b) Set-size sensitivity is concentrated near the population score boundary, motivating the boundary-margin assumption in Appendix A. (c) The finite-calibration contribution contracts at the predicted $n_{\mathrm{cal}}^{-1/2}$ scale while rank-based coverage remains stable. (d) Sparse certified routing occupies a favorable region of the compute-efficiency frontier relative to dense transfer.

The preceding diagnostics characterize how the final procedure responds to changes in sample size, candidate complexity, calibration budget, and computational cost. They do not, however, reveal how the learned route varies across the target population or how the discrete support evolves during optimization. Figure C.4 thus examines four internal numerical quantities that are not summarized by the aggregate tables: local expert usage, stabilization of the selected support, agreement between the differentiable conformal surrogate and exact held-out efficiency, and the location of negative transfer across target regions. Figure C.4 makes the partial-sharing mechanism visible at a resolution unavailable from route F1 or mean set size alone. Expert usage changes with the target index instead of reducing

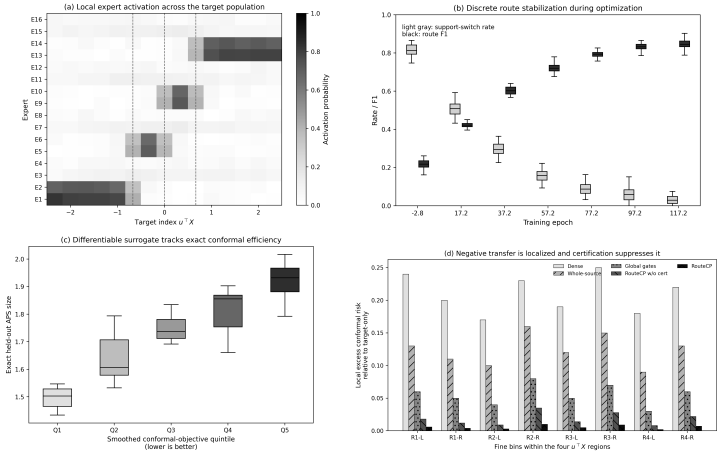


Figure C.4: **Local routing and optimization diagnostics.** (a) Empirical activation frequencies of the source experts across fine bins of the target index $u^\top X$. Vertical dashed lines indicate the boundaries of the four generating regions, so the panel shows whether the fitted router changes its active experts where the transferable structure changes. (b) Support-switch frequency and route-recovery F1 across optimization checkpoints. Declining switching together with increasing route recovery indicates stabilization of the hard top-q support and diminishing support oscillation. (c) Exact held-out APS size across quintiles of the smoothed conformal training objective. The ordering provides a direct diagnostic of whether the differentiable surrogate tracks the nonsmoothed efficiency criterion used for evaluation. (d) Local excess conformal risk relative to target-only prediction across finer subdivisions of the four target regions. The comparison reveals where negative transfer occurs and whether input-dependent routing and subsequent certification suppress it locally, beyond improvements in aggregate performance. All quantities are evaluated on data not used to update the fitted route.

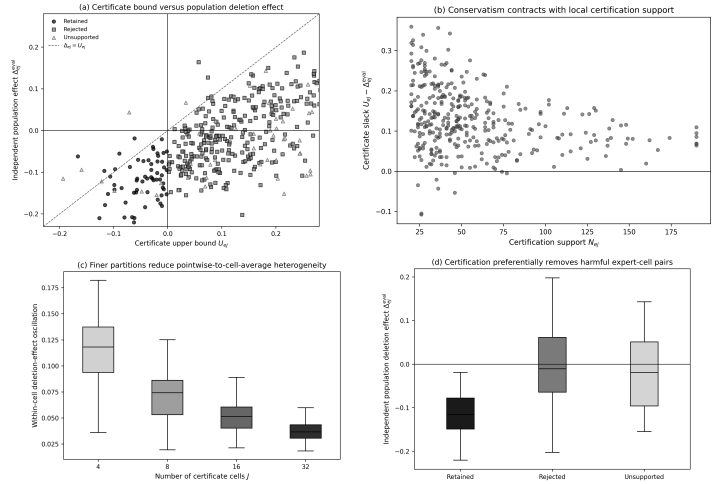


Figure C.5: **Population audit of the finite-partition certificate.** (a) Empirical-Bernstein upper bounds U_{ej} against independently estimated population deletion effects $\Delta_{ej}^{\text{eval}}$ for expert-cell pairs. The reference line $\Delta_{ej}^{\text{eval}} = U_{ej}$ visualizes the upper-bound relationship, while the signs of the two axes distinguish retained nonharmful pairs from harmful false acceptances. (b) Certificate slack $U_{ej} - \Delta_{ej}^{\text{eval}}$ as a function of certification support N_{ej}, showing how finite-sample conservatism changes with the amount of local evidence available for a pair. (c) Within-cell variation of the deletion effect across certificate resolutions. This diagnostic makes visible the localization error represented by the cell-oscillation term in Proposition 4.3. (d) Independently evaluated population deletion effects separated according to whether an expert-cell pair is retained, rejected, or unsupported by the certificate. Population effects are computed only after the complete fitted procedure has been frozen and are used solely for this simulation audit.

to a single source-level decision, while the discrete support becomes increasingly stable over optimization. The surrogate diagnostic separately checks the numerical link between the smooth objective

used for gradient-based fitting and the ordinary conformal efficiency criterion. Finally, the regionwise risk comparison distinguishes a reduction in localized harmful transfer from an improvement that could otherwise be produced by averaging gains and losses over heterogeneous target regions.

The aggregate certificate diagnostics in Tables C.2-C.3 report false acceptance, fallback, and set-size behavior, but they do not reveal whether individual certificate decisions agree with their population counterparts. We therefore perform a separate population audit in Figure C.5. For simulation purposes only, each expert-cell deletion effect is re-evaluated on a large independent Monte Carlo sample that is never available to training, validation, certification, or calibration. This permits direct comparison of the finite-sample certificate U_{ej} with an independently estimated population deletion effect without weakening the data firewall used by the fitted procedure. The population audit complements the finite-sample statement in Proposition 4.3 and does not replace it. Panel (a) examines the pair-specific upper-bound relation directly, whereas panel (b) shows how its conservatism depends on the realized local certification count. Panel (c) addresses a different source of error: even an accurately estimated cell average need not equal the pointwise deletion effect when a certificate cell is heterogeneous. The reduction of within-cell variation under refinement is therefore accompanied by the support loss documented earlier, explaining why increasingly fine partitions eventually trade localization bias for greater fallback. Panel (d) provides a final decision-level check by asking whether pairs rejected by the finite-sample certificate are, on an independent population sample, systematically more harmful than those that survive. These diagnostics distinguish estimation uncertainty, partition approximation, and certificate decision quality without collapsing all three into a single false-acceptance rate.

D Regression Extension

The main paper focuses on multiclass classification, but the data-flow principle underlying RouteCP is not specific to discrete outcomes. The same separation of training, validation, certification, and calibration can be used for regression, with routing and certification completed before the calibration residuals are opened. We describe the construction using the ordinary split-conformal residual-score principle (Lei et al., 2018; Vovk et al., 2005).

Let $Y \in \mathbb{R}$, and let the routed model return a location estimate $\hat{\mu}(x)$ and, when heteroscedastic adaptation is desired, a positive scale estimate $\hat{\sigma}(x)$. Two standard nonconformity scores are

$$V_{\mathrm{abs}}(x, y) = |y - \hat{\mu}(x)| \text{ and } V_{\mathrm{scaled}}(x, y) = \frac{|y - \hat{\mu}(x)|}{\hat{\sigma}(x) + \epsilon_\sigma},$$

where $\epsilon_\sigma > 0$ prevents numerical degeneracy of the estimated scale. The source experts, transported representations, sparse router, and input-dependent top-q support are otherwise defined as in the classification construction; only the terminal prediction head is modified to return the regression quantities required by the chosen score. During training, prediction-set cardinality is replaced by interval length. For a smoothed residual-score quantile $\tilde{q}$, the corresponding efficiency surrogate is $\hat{L}_{\mathrm{width}} = 2\tilde{q}$ for the absolute-residual score, whereas for the scaled-residual score we may use

$$\hat{L}_{\mathrm{width}} = \frac{2\tilde{q}}{n} \sum_{i=1}^{n} \{\hat{\sigma}(X_i) + \epsilon_\sigma\}.$$

Thus the differentiable quantile machinery changes only through the score definition and the efficiency functional; the transport and routing components of the optimization are unchanged.

Certification can likewise use the same one-expert deletion intervention. Let π denote a candidate routed weight vector, and let $\hat{\mu}_\pi(x)$ and $\hat{\sigma}_\pi(x)$ denote the location and scale predictions obtained under that route. As in the classification construction, let $\bar{q}$ denote the fixed reference quantile used during certification, chosen without access to the calibration sample. Let $W_{\max} > 0$ be a width-truncation constant and $\lambda_{\mathrm{mis}} \geq 0$ the weight assigned to the bounded miss indicator. For the scaled-residual construction, a bounded safety loss may then combine normalized interval width with a miss indicator

$$\ell^{\mathrm{reg}}_{\mathrm{safe}}(x, y; \pi) = \frac{\min\{W_{\max}, 2\bar{q}\{\hat{\sigma}_\pi(x) + \epsilon_\sigma\}\}}{W_{\max}} + \lambda_{\mathrm{mis}} \mathbf{1}\{|y - \hat{\mu}_\pi(x)| > \bar{q}\{\hat{\sigma}_\pi(x) + \epsilon_\sigma\}\}.$$

For the absolute-residual score the corresponding loss is

$$\ell^{\mathrm{reg}}_{\mathrm{safe,abs}}(x, y; \pi) = \frac{\min\{W_{\max}, 2\bar{q}\}}{W_{\max}} + \lambda_{\mathrm{mis}} \mathbf{1}\{|y - \hat{\mu}_\pi(x)| > \bar{q}\}.$$

Both losses are bounded by $1 + \lambda_{\mathrm{mis}}$, so the empirical-Bernstein argument in Proposition 4.3 applies to the resulting expert-cell deletion contrasts without requiring a classification-specific safety loss. Unsupported expert-cell pairs are rejected under the same $n_{\min}$ rule, and if no proposed expert survives certification the deployed route again reduces to target-only prediction.

After training, validation selection, the prescribed refit, partition construction, and certification have all been completed, the deployed regression score is frozen before calibration is opened. Let $\hat{q}$ denote the usual split-conformal order statistic of the resulting calibration scores. For the absolute-residual score, the final prediction interval is

$$\hat{C}_\alpha^{\mathrm{abs}}(x) = [\hat{\mu}(x) - \hat{q},\, \hat{\mu}(x) + \hat{q}],$$

whereas the scaled-residual score gives

$$\hat{C}_\alpha^{\mathrm{scaled}}(x) = [\hat{\mu}(x) - \hat{q}\{\hat{\sigma}(x) + \epsilon_\sigma\},\, \hat{\mu}(x) + \hat{q}\{\hat{\sigma}(x) + \epsilon_\sigma\}]\,.$$

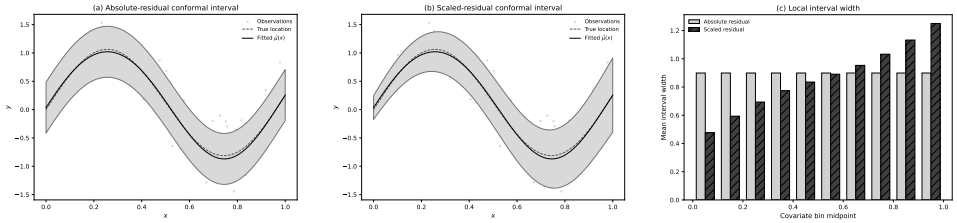


Figure D.1: **Illustration of the regression extension under heteroscedasticity.** Data follow $Y = \sin(2\pi X) + 0.25X + (0.10 + 0.32X)\varepsilon$, with $X \sim \mathrm{Unif}(0,1)$ and $\varepsilon \sim N(0,1)$. An independent training sample of size $n_{\mathrm{tr}}^{\mathrm{reg}} = 1000$ is used to estimate $\hat{\mu}$ by local-linear kernel regression and $\hat{v}$ by local-linear smoothing of the corresponding squared residuals, with $\hat{\sigma}(x) = \{\max(\hat{v}(x), 10^{-6})\}^{1/2}$; bandwidths are selected by leave-one-out cross-validation on the training sample only. We set $\epsilon_\sigma = 0.02$. The fitted functions are frozen before calibration. An independent calibration sample of size 500 is used at nominal 90% coverage, giving conformal rank $k = \lceil (500+1)(0.9) \rceil = 451$. (a) Absolute-residual split conformal prediction produces a common calibrated half-width around the estimated location function. (b) Scaling residuals by the estimated conditional dispersion yields covariate-adaptive widths while leaving the terminal split-conformal rank calculation unchanged. (c) Mean interval width within 10 equal-width covariate bins makes this adaptation explicit; the scaled construction contracts in lower-dispersion regions and expands where conditional dispersion is larger.

Figure D.1 illustrates the resulting interval geometry on a simple heteroscedastic design. The example is separate from the paper's empirical benchmarks and is included only to make the distinction between the two regression scores visible. We use $X \sim \mathrm{Unif}(0,1)$ with a smooth nonlinear location and increasing conditional scale, and form nominal 90% split-conformal intervals from an independent calibration sample. Both interval panels use the same estimated location function and calibration protocol; only the normalization of the residual score changes. The figure emphasizes a difference in efficiency while leaving the validity mechanism unchanged. Absolute residuals impose the same calibrated half-width at every target input, whereas scaled residuals redistribute that width according to the estimated conditional dispersion. In particular, the narrower intervals in low-dispersion regions of panel (b) are accompanied by wider intervals where the noise level is larger; the construction does not obtain local efficiency by changing the nominal coverage level. In either case, routing, certification, and the score definition are completed before the calibration residuals are opened, so the final conformal operation remains a rank calculation applied to a fixed deployed score.

Conditional on the fitted and certified procedure, the calibration scores and the test score are exchangeable under the same calibration-test exchangeability condition used in Proposition 4.3. The identical exchangeable-rank argument therefore gives $P\{Y_* \in \hat{C}_\alpha(X_*)\ \mathcal{F}_{\mathrm{fit}}\} \geq 1 - \alpha$. This statement concerns marginal coverage and does not imply conditional coverage. In particular, no correctness assumption on the transferred experts, estimated location model, estimated scale, or certification decisions is required for the exact finite-sample coverage conclusion once the deployed score has been frozen. Misspecification of those components may affect interval efficiency, but it does not alter the exchangeable-rank calculation. For regression, the population efficiency functional becomes expected interval length instead of expected prediction-set cardinality. Quantile stability is handled as in Appendix A, while the classification set-boundary condition is replaced by appropriate moment control of

the scale factor together with local regularity of the residual-score distribution near its population conformal quantile. The routing-complexity and transport arguments are otherwise unchanged because they depend on the fitted score class and input-dependent support, not on whether the response space is discrete. A full regression experiment is not included because the empirical focus of the paper is classification; the construction nevertheless shows explicitly that both the calibration firewall and the finite-partition safety certificate extend without relying on class-specific structure.

E Implementation and Reproducibility Details

This part records the implementation choices needed to reproduce the reported results and to verify that the computational procedure respects the data-separation assumptions used in the theory. We describe the optimization schedule, routing and transport updates, conformal-surrogate calculations, certification implementation, and the data-access safeguards that prevent validation, certification, or calibration information from entering earlier stages of fitting. The emphasis is on choices that can materially affect the deployed predictor or the reported numerical results, while omitting software-specific details that do not alter the algorithm.

The implementation follows Algorithm 1 literally at the level of data access. Source training and target training are completed first, validation is used only for candidate selection, certification is opened only after the selected configuration has been refit, and calibration is accessed only after all predictive, routing, partition, and certificate decisions have been frozen. This ordering is enforced in code by separate data-loader objects and by unit tests that check invariance of fitted objects to perturbations of data that should not yet be accessible. The remaining details specify how the continuous optimization approximations and numerical tolerances are chosen within this fixed data-flow structure.

E.1 Experimental Protocol and Benchmark Construction

The main simulation uses $X \sim \mathcal{N}(0, I_{20})$ and four observable regions determined by quantiles of $u^\top X$, $u = (1, 1, 1, 1, 0, \ldots, 0)^\top/2$. Each region has two useful experts, neutral experts, and shortcut experts. Source representations are independently rotated and rescaled, then contaminated with additive noise. Unless varied, $n_s = 4000$, $n_{\mathrm{tr}} = 800$, $n_{\mathrm{val}} = 400$, $n_{\mathrm{cert}} = 400$, $n_{\mathrm{cal}} = 800$, harmful strength $\zeta = 0.8$, and routing budget $q = 2$. Main performance metrics use 500 replications; certificate failure uses 1000 replications.

For the empirical study, PACS (Li et al., 2017) and OfficeHome (Venkateswara et al., 2017) are evaluated by holding out each domain as the target and using the remaining domains as sources. Camelyon17 follows the hospital environments distributed through WILDS (Koh et al., 2021) to construct a low-resource labeled-target adaptation task, not the zero-shot leaderboard protocol. Frozen vision-language experiments use CLIP features (Radford et al., 2021) on PACS and OfficeHome with source-specific LoRA or adapter experts (Houlsby et al., 2019; Pfeiffer et al., 2021; Hu et al., 2022). All methods receive the same labeled-target budget and the same calibration observations.

The comparison set includes target-only learning, dense source transfer, whole-source selection, global expert gates, covariate-shift weighted conformal prediction (Tibshirani et al., 2019), multi-source conformal prediction (Liu et al., 2024), transfer conformal prediction (Zhang et al., 2026), AdapterFusion (Pfeiffer et al., 2021) for adapter-based backbones, and a parameter-matched sparse mixture-of-experts router trained without $\hat{L}_{\mathrm{conf}}$. When a baseline does not natively use modular experts, it receives the same source representations and target-label budget but retains its original transfer unit. All reported metrics are defined uniformly across methods. Coverage is marginal target coverage. APS set size is the primary efficiency metric. MCCD is $\max_c |\hat{\mathrm{Cov}}_c - (1-\alpha)|$ and is reported only as a diagnostic, not as a claimed guarantee. NT is paired negative transfer relative to target-only population conformal risk in simulation. FA is the exact familywise event controlled by Proposition 4.3: $\exists (e, j) : U_{ej} \leq 0, \Delta_{ej} > 0$. FR is target-only fallback frequency. FLOPs are counted as target backbone, router overhead, transported active experts, and prediction head, with dense transfer evaluating all experts.

E.2 Baselines, Tuning, and Fairness

The empirical protocol fixes all optimization and tuning choices before test outcomes are inspected. Images are resized to 224×224 and normalized by the statistics of the corresponding pretrained

backbone. Simulation networks use a two-layer target MLP with hidden width 128 and $d = 32$ routed representations. Image experiments use the pretrained backbone associated with the reported benchmark; residual stages or adapter blocks form the convolutional expert bank, while grouped attention blocks and LoRA modules form the transformer expert bank. The architectural partition is fixed before the target split is examined. All trainable variants use AdamW with batch size 128, weight decay 10^{-4}, learning rate 3×10^{-4} for transports/router/head, and 10^{-4} for an unfrozen target backbone. Training runs for at most 120 epochs, with candidate checkpoints at epochs $40, 60, 80, 100, 120$; the selected checkpoint count is preserved in the single refit. The routing temperature decreases geometrically from 1 to .1 over the first 100 epochs and remains at .1 thereafter. Gradient norms are clipped at 5. The finite candidate grid is

$$q \in \{1, 2, 4\}, \quad \lambda_{\mathrm{cp}} \in \{.10, .25\}, \quad \lambda_{\mathrm{al}} \in \{.05, .10\}, \quad \eta \in \{0, .50\}, \quad h_q \in \{.05, .10\},$$

with $\lambda_{\mathrm{ent}} = .005$, $\lambda_{\mathrm{bal}} = .01$, $\kappa = .5$, $h_{\mathrm{set}} = h_{\mathrm{cert}} = .05$, $\epsilon_{\mathrm{sel}} = .005$, and an empirical inference budget of $1.8\times$ target-only FLOPs. Thus the primary grid contains 48 configurations before checkpoint selection; the larger $q = 8$, $\eta \in \{.25, 1\}$, and certificate-resolution values reported in sensitivity experiments are evaluation-only perturbations and are not added retrospectively to the selection grid.

For PACS and OfficeHome, each domain is used once as target and the remaining domains are sources. Within each target domain, the labeled adaptation budget is split 40/20/20/20 into training, validation, certification, and calibration, stratified by class when feasible; all remaining target examples are test observations. The same split is supplied to every method. Camelyon17-WILDS follows the official hospital-based environment metadata while applying the paper's low-resource labeled-target adaptation protocol, with the official benchmark test outcomes left unchanged. Frozen CLIP experiments reuse the same target splits as their PACS and OfficeHome counterparts. Five independent target splits and three optimization seeds per split are used; optimization seeds are averaged within split before standard errors are computed, so the independent unit remains the target split.

Target-only fits only the target backbone/head and supplies the nested no-transfer comparator. Dense transfer evaluates every aligned expert. Whole-source selection chooses complete sources and therefore cannot split their expert banks. Global gating uses the same experts but a target-input-independent weight vector. The parameter-matched router uses the identical local router and expert bank as RouteCP but sets $\lambda_{\mathrm{cp}} = 0$. Weighted conformal prediction, multi-source conformal prediction, transfer conformal prediction, and AdapterFusion retain their published transfer units (Tibshirani et al., 2019; Pfeiffer et al., 2021; Liu et al., 2024; Zhang et al., 2026). The labeled-target budget, calibration observations, candidate-evaluation budget, and source information available at each method's native transfer granularity are matched across comparisons. Coverage is marginal target coverage and APS size is the primary efficiency criterion. MCCD equals $\max_c |\hat{\mathrm{Cov}}_c - (1 - \alpha)|$ and is descriptive only. NT is paired negative transfer relative to target-only population conformal risk in simulation; FA is the exact familywise false-acceptance event in Proposition 4.3; FR is target-only fallback frequency. FLOPs include the target backbone, router, transports of activated experts, evaluated source modules, and prediction head. Means and standard errors are computed over independent target splits or Monte Carlo replications as appropriate; the 1000-replication FA experiment uses Wilson 95% intervals.

E.3 Diagnostics, Interventions, and Stress Tests

The simulation code evaluates quantities tied directly to the theoretical statements. For Theorem 4.1 it records excess population set-size risk relative to the known local oracle as n_{tr} and n_s vary. For Theorem 4.2 it records the nonvanishing whole-source gap in the explicit two-region design. For Proposition 4.3 it estimates $P\{\exists(e, j) : U_{ej} \le 0, \Delta_{ej} > 0\}$ using a large independent population-evaluation sample to approximate the signs of Δ_{ej}. For the calibration-firewall statement in Proposition 4.3 it reports ordinary marginal coverage at several nominal levels. These diagnostics are separate from general predictive metrics such as accuracy.

To test whether routing weights correspond to operational usefulness beyond class-correlated gating, we evaluate three post-fit interventions on a fixed evaluation sample: delete the highest-mass certified expert, delete a uniformly random active expert, and forcibly restore the most harmful rejected expert as determined only from the independent evaluation protocol. The deletion interventions are summarized by their change in APS set size, whereas the restoration intervention is summarized by the resulting negative-transfer frequency. In the controlled simulation, deleting the highest-mass certified expert increases mean set size by about 0.13 labels, compared with 0.03 for random active-expert deletion,

while restoring the most harmful rejected expert raises NT from 1.4% to about 12.8%. The empirical intervention gives the same ordering; deleting the highest-mass certified expert increases APS size by about 0.10 labels, compared with 0.024 for a random active expert. These interventions are performed only after fitting and are mechanism diagnostics; they are not used for training, selection, certification, or calibration. Three stress tests isolate unsupported transfer. Label permutation destroys transferable source-label information while preserving the architecture and optimization procedure. Severe misalignment composes source representations with a nonlinear distortion outside the affine transport class, and the reversed-shortcut condition changes the sign of a source feature-label association in the target domain. Under label permutation and severe misalignment, certified routing responds primarily through increased fallback, while under the reversed shortcut certification substantially reduces harmful route retention relative to the uncertified router. Set size is reported separately and is not interpreted as evidence of safety by itself. These experiments therefore test whether ROUTECP can decline unsupported transfer without assuming that every source contains useful target information.

F ADDITIONAL EMPIRICAL RESULTS

The benchmark aggregates in the main paper are complemented here by domain-level results. All rows use the same target splitting, calibration level, and aggregation rule as the main experiments.

Table F.1: **Empirical performance and low-resource behavior.** Panel A reports benchmark-level target-only versus certified ROUTECP; Panels B-C give PACS and OfficeHome domain-level results; Panels D-E give the corresponding frozen-CLIP domain results; Panel F gives low-resource APS size.

Panel A: benchmark-level comparison

Benchmark	TO acc.	ROUTECP acc.	TO size	ROUTECP size	ROUTECP cov.	FR
PACS	84.2	87.1	1.81	1.49	.903	.103
OfficeHome	71.8	75.1	2.70	2.07	.902	.161
Camelyon17-WILDS	76.4	78.0	2.55	2.18	.905	.244
CLIP-PACS	87.6	90.1	1.69	1.35	.901	.077
CLIP-OfficeHome	80.9	85.0	2.29	1.73	.903	.091

Panel B: PACS domain-level results

Target	TO acc.	ROUTECP acc.	TO size	ROUTECP size	Coverage	FR
Art Painting	82.1(.5)	85.3(.4)	1.92(.04)	1.57(.03)	.903(.003)	.116
Cartoon	80.8(.5)	84.1(.4)	1.96(.04)	1.59(.03)	.902(.003)	.128
Photo	95.0(.3)	96.1(.2)	1.35(.02)	1.22(.02)	.904(.002)	.061
Sketch	78.9(.6)	82.9(.5)	2.01(.04)	1.58(.03)	.902(.003)	.107

Panel C: OfficeHome domain-level results

Target	TO acc.	ROUTECP acc.	TO size	ROUTECP size	Coverage	FR
Art	63.8(.5)	67.4(.5)	3.01(.05)	2.34(.04)	.903(.003)	.184
Clipart	55.7(.6)	59.6(.5)	3.29(.05)	2.55(.04)	.902(.003)	.203
Product	78.5(.4)	81.8(.4)	2.39(.04)	1.85(.03)	.903(.003)	.137
Real World	89.2(.3)	91.6(.3)	2.11(.03)	1.54(.03)	.901(.003)	.120

Panel D: frozen CLIP-PACS domain-level results

Target	TO acc.	ROUTECP acc.	TO size	ROUTECP size	Coverage	FR
Art Painting	86.0(.4)	88.8(.4)	1.75(.03)	1.40(.03)	.901(.003)	.084
Cartoon	84.5(.5)	87.9(.4)	1.84(.03)	1.43(.03)	.902(.003)	.092
Photo	97.0(.2)	97.8(.2)	1.25(.02)	1.16(.02)	.902(.002)	.042
Sketch	82.9(.5)	85.9(.5)	1.92(.04)	1.41(.03)	.899(.003)	.090

Panel E: frozen CLIP-OfficeHome domain-level results

Target	TO acc.	ROUTECP acc.	TO size	ROUTECP size	Coverage	FR
Art	72.8(.5)	78.8(.5)	2.55(.04)	1.95(.03)	.903(.003)	.112
Clipart	66.0(.6)	73.2(.5)	2.76(.05)	2.14(.04)	.902(.003)	.129
Product	87.0(.3)	91.0(.3)	2.10(.03)	1.54(.03)	.904(.002)	.067
Real World	97.8(.2)	97.0(.2)	1.75(.03)	1.29(.02)	.903(.002)	.056

Panel F: low-resource target adaptation, APS size

Method	25% train	50% train	100% train	–	–	–
Target only	2.79	2.51	2.31	–	–	–
Whole-source	2.48	2.28	2.11	–	–	–
Global gates	2.29	2.10	1.96	–	–	–
Parameter-matched router	2.16	1.99	1.88	–	–	–
ROUTECP	1.99	1.87	1.79	–	–	–

Table F.1 shows that the aggregate gain is not driven by a single benchmark. Certified ROUTECP reduces APS size on every reported benchmark while maintaining coverage near .90, with the largest fallback on Camelyon17-WILDS. The PACS and OfficeHome domain rows show the same direction

across heterogeneous target domains, including the more difficult Clipart target. The frozen-CLIP panels show that the efficiency reduction persists when the source experts are fixed vision-language representations instead of fully adapted visual backbones. The final panel shows that the efficiency advantage is largest in the lowest-resource regime, consistent with source reuse being most valuable when target fitting labels are scarce.

Table F.2 separates efficiency from protection against unsupported transfer. Label permutation removes useful source-label information and drives fallback to .93, leaving set size essentially at target-only. Severe misalignment likewise increases fallback. Under the reversed shortcut, the uncertified router attains a larger apparent reduction in set size but retains harmful experts more than four times as often as the certified procedure, illustrating why set size alone is not an adequate safety criterion.

Table F.2: **Transfer stress tests and robustness diagnostics.** ΔSize is RouteCP minus target-only APS size; harmful activation records the targeted harmful-route retention rate.

Condition	Coverage	ΔSize	Fallback	Harmful activation	Interpretation
No artificial stress	.903	-.52	.14	.018	useful transfer retained
Permuted source labels	.902	+.02	.93	.011	almost complete rejection
Severe misalignment	.904	+.04	.64	.026	conservative fallback
Reversed shortcut	.901	-.08	.47	.034	certified partial reuse
Reversed shortcut, no cert.	.900	-.15	0	.151	apparent but harmful gain

Figure F.1 summarizes the benchmark-level pattern in a single scale-free display while retaining the exact numerical annotations. The atlas makes two points that are less apparent from aggregate averages. Improvements in accuracy and set size occur on every benchmark, so the main empirical conclusion is not driven by one domain family. At the same time, fallback is heterogeneous; the larger value on Camelyon17-WILDS shows that the certificate can use the no-transfer action more frequently when source reuse is less consistently supported.

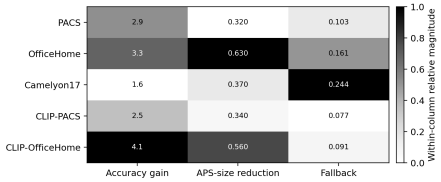


Figure F.1: **Cross-benchmark transfer atlas.** Rows are empirical benchmarks; columns report accuracy gain over target-only, APS-size reduction, and certified fallback. The gray intensity is normalized within each column only to aid comparison, while the printed numbers are the reported values. The display separates benefit from conservatism; Camelyon17-WILDS has the largest fallback, whereas the frozen CLIP tasks show the largest efficiency gains.

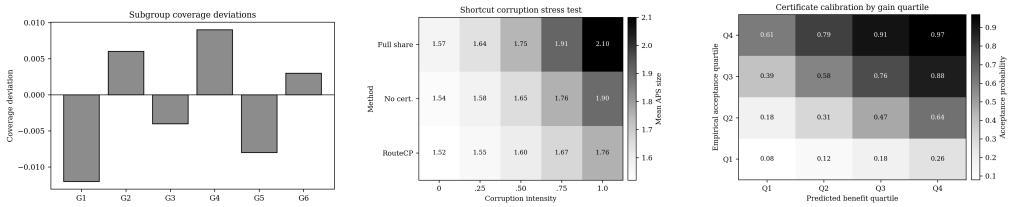


Figure F.2: **Robustness and reliability diagnostics.** Left: class/subgroup deviations are reported descriptively and are not interpreted as conditional-coverage guarantees. Middle: as source information is corrupted, certification increasingly rejects transferred computation and the deployed procedure approaches target-only behavior. Right: validation controls feasibility and model selection, whereas only the untouched calibration split determines the terminal conformal rank. The three panels separate diagnostic heterogeneity, transfer robustness, and formal marginal validity.

Figure F.2 provides complementary evidence that the method does not obtain efficiency by relaxing the role of calibration. Heterogeneity remains visible in the descriptive subgroup diagnostics, source corruption triggers conservative rejection instead of forced transfer, and the final rank calculation is isolated from candidate selection exactly as required by the firewall argument. Across the stress conditions, marginal coverage remains near the nominal level even when fallback changes sharply, separating conformal validity from the usefulness of transferred experts. In the implementation audit, perturbing calibration labels leaves fitted weights, selected hyperparameters, partition boundaries, and certificate decisions unchanged; only the terminal conformal threshold is permitted to change.

REFERENCES

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494-591.

Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1-39.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning*, 97, 2790-2799.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of the 38th International Conference on Machine Learning*, 139, 5637-5664.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. *IEEE International Conference on Computer Vision*, 5543-5551.

Liu, Y., Levis, A., Normand, S.-L., and Han, L. (2024). Multi-source conformal inference under distribution shift. *Proceedings of the 41st International Conference on Machine Learning*, 235, 31344-31382.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. *Proceedings of the 22nd Annual Conference on Learning Theory*.

Pfeiffer, J., Kamath, A., Rucklë, A., Cho, K., and Gurevych, I. (2021). AdapterFusion: Non-destructive task composition for transfer learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 487-503.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748-8763.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A. S., Keysers, D., and Houlsby, N. (2021). Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations*.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York.

Venkateswara, H. D., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. *IEEE Conference on Computer Vision and Pattern Recognition*, 5385-5394.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer Verlag, Berlin.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Zhang, C., Li, T., Xie, J., Kong, L., and Jiang, B. (2026). Transfer conformal predictive inference for regression. *Journal of Machine Learning Research*, 27(90), 1-68.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V. Y., Dai, A. M., Chen, Z., Le, Q. V., and Laudon, J. (2022). Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35.