ADAPTIVE CONFORMAL PREDICTION VIA LEARNED RESIDUAL-LAW RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction provides finite-sample marginal coverage, but adapting prediction sets to query-dependent multivariate uncertainty remains difficult in high dimensions. We introduce Factorized Conformal Prediction (FacCP), which estimates a query-specific residual law and separates uncertainty learning from conformal calibration. Rather than directly learning a score, quantile, or final set, FacCP learns an uncertainty-relevant representation for retrieving residuals with similar conditional laws; the retrieved law determines local geometry, while a single global conformal quantile calibrates its inflation. We formalize a geometry-approximation condition under which error in local residual geometry directly bounds score non-pivotality, linking geometry quality to local coverage under a single global conformal calibration. The retrieved residual law can instantiate multiple set families without retraining. Experiments validate the connection between residual-law recovery, adaptive geometry, and pivotality, and show that local calibration alone cannot recover misspecified geometry. Across high-dimensional robotic reachability and autonomous driving, FacCP yields tighter calibrated sets, with particularly strong data and parameter efficiency in low-data and low-capacity regimes.

1 INTRODUCTION

Uncertainty quantification (UQ) constructs high-probability prediction sets for reliable downstream decision-making. These sets should achieve *reliable coverage* while remaining *tight*, even with limited data. This requires uncertainty estimates that are both *globally valid* and *locally tailored* to each query. Overly conservative sets can, for example, eliminate feasible maneuvers in autonomous driving or make robust control inefficient. This adaptivity is difficult when uncertainty varies across operating conditions and local data is scarce, especially in high-dimensional input spaces where naïve localization to nearby examples can further reduce effective sample size. We thus seek prediction sets that adapt to query-dependent uncertainty while making effective use of limited data.

Conformal prediction (CP) is appealing for UQ since it converts a fixed predictive score into finite-sample marginal coverage under exchangeability. However, marginal validity alone does not ensure locally appropriate or geometrically efficient sets. When residual distributions vary across queries, e.g., in orientation or higher moments, a single fixed global geometry can be overly conservative. Existing adaptive CP methods consider this issue. Learned and normalized scores adapt a *scalar* notion of difficulty (Romano et al., 2019; Seedat et al., 2023; Laplante, 2026; Plassier et al., 2025a), but remain limited by the geometry of the chosen score. Localized CP estimates query-dependent quantiles (Guan, 2023; Hore & Barber, 2025; Kiyani et al., 2024), but this reduces effective calibration support when local data is scarce, leading to noisy estimates. Multivariate methods learn adaptive covariances, transformations, or densities (Messoudi et al., 2022; Fang et al., 2025; English & Lippert, 2026), enabling richer geometry but coupling learning to a particular uncertainty model and set family. Overall, existing methods either restrict adaptation to a fixed scalar geometry, sacrifice calibration data support via localization, or are restricted to a task-specific set family.

To close these gaps, we introduce *Factorized Conformal Prediction (FacCP)*. Rather than directly learning the final set or a local conformal quantile, FacCP learns which observed residuals are informative about a query's conditional residual law. An energy-score objective trains an uncertainty-relevant representation in which proximity reflects similarity of conditional residual distributions. The retrieved empirical law determines a query-dependent geometry, such as an ellipsoid or hyperbox, which is then reduced to a scalar nonconformity score. A single split-conformal threshold

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.

is estimated using the *entire* calibration sample. Thus, *local* data determines uncertainty *shape*, while *global* calibration determines only its final *inflation*. FacCP preserves exact finite-sample marginal validity while targeting the full conditional residual law. We characterize uncertainty-sufficient representations, show that removing nuisance variation improves localization efficiency, and establish a geometry-approximation condition under which local geometry error bounds score non-pivotality, yielding conditional-coverage bounds separating geometry and calibration error. Approximate pivotality also supports covariate-shift transfer. Experiments on controlled and high-dimensional robotic tasks show calibrated, efficient sets that remain robust in data- and capacity-limited regimes. Our specific contributions are:

1. We introduce *Factorized Conformal Prediction*, which separates query-dependent residual-law estimation and geometry construction from a single globally pooled conformal calibration step, enabling adaptive multivariate prediction sets without fragmenting the calibration sample.
2. We develop an energy-score objective for learning an uncertainty-relevant retrieval representation whose population target is the full conditional residual law, and characterize uncertainty sufficiency and the statistical benefit of suppressing uncertainty-irrelevant dimensions.
3. We theoretically analyze FacCP, connecting residual-geometry approximation to approximate score pivotality and, through pivotality, to local coverage under global calibration, separating the roles of representation/localization error, geometry approximation, and finite calibration.
4. We empirically validate FacCP on high-dimensional robotic reachability tasks and autonomous driving UQ, demonstrating robust set efficiency across data- and model-capacity limitations.

2 Related Work

CP provides distribution-free marginal coverage under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018), while exact distribution-free conditional coverage is generally impossible without additional assumptions (Lei & Wasserman, 2014; Barber et al., 2021). This motivates scores whose distributions are more comparable across queries. In classification, Romano et al. (2020) construct an oracle conditional-mass score that is uniform across queries. CQR and learned or normalized scores adapt to heteroscedasticity or instance difficulty (Romano et al., 2019; Seedat et al., 2023; Amoukou & Brunel, 2023; Colombo, 2024; Deutschmann et al., 2024), while score-rectification methods target conditional score homogeneity (Laplante, 2026; Plassier et al., 2025a). Such scalar adaptation, however, remains tied to the geometry of the underlying score and cannot recover multivariate residual structure absent from it. FacCP instead targets the conditional residual law before scalarization, so geometry itself can adapt to the query.

Localized CP estimates query-dependent quantiles or modifies scores using local conditional information (Guan, 2023; Hore & Barber, 2025; Kiyani et al., 2024; Gibbs et al., 2025; Han et al., 2022). These approaches address score heterogeneity, but localization can reduce effective support and make tail estimation noisy under limited data. FacCP moves localization upstream to residual-law retrieval while retaining a single globally calibrated quantile, preserving calibration support and suppressing uncertainty-irrelevant variation. Other methods construct adaptive sets from conditional distributions, densities, or generative models (Lei & Wasserman, 2014; Chernozhukov et al., 2021; Izbicki et al., 2020; 2022; Diamant et al., 2024; Wang et al., 2023; Plassier et al., 2025b), or directly learn multivariate ellipsoids, representations, templates, flows, or transport maps (Messoudi et al., 2022; Xu et al., 2024; Feldman et al., 2023; Tumu et al., 2024; Dheur et al., 2025; Thurin et al., 2025; Fang et al., 2025; English & Lippert, 2026). These approaches can represent rich geometry, but generally require learning a particular conditional uncertainty model or couple the learned representation to the final set family. FacCP instead retrieves observed residuals in an uncertainty-relevant latent space and separates residual-law representation from geometry construction, allowing the same representation to support multiple set families without retraining. Weighted CP handles covariate shift through reweighted calibration (Tibshirani et al., 2019; Barber et al., 2023; Joshi et al., 2026). FacCP is complementary: approximate pivotality allows the same unweighted global threshold to transfer across shifts preserving the conditional response law, although shift-aware calibration remains necessary when pivotality fails. Related adaptive conformal geometries have also been used for robust reachable tubes (Srinivasan et al., 2026). App. B provides more related work discussion.

3 Problem Statement

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the query and response spaces, and let $P_{XY} : \mathcal{B}(\mathcal{X} \times \mathcal{Y}) \to [0, 1]$ be the joint data-generating distribution of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, with $(X, Y) \sim P_{XY}$. We are given a fixed predictor

$\widehat{f} : \mathcal{X} \to \mathcal{Y}$ and a task-specific measurable residual map $T : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$, where $\mathcal{R} \subseteq \mathbb{R}^{d_R}$. Define

$$R := T(X, Y), \qquad P_x^R := \mathcal{L}(R \mid X = x), \tag{1}$$

where $\mathcal{L}(R \mid X = x)$ denotes the conditional law of R given $X = x$. Equivalently, P_x^R is a probability measure on $\mathcal{R}$ satisfying $P_x^R(A) = \mathbb{P}(T(X, Y) \in A \mid X = x)$ for every measurable $A \subseteq \mathcal{R}$. The residual may be scalar or vector-valued. We focus on settings in which the conditional residual law P_x^R is heterogeneous across x, potentially with large variations in mean or higher moments.

We are given two disjoint datasets: $\mathcal{D}_{\text{ref}} = \{(X_j^{\text{ref}}, Y_j^{\text{ref}})\}_{j=1}^m$ and $\mathcal{D}_{\text{cal}} = \{(X_i^{\text{cal}}, Y_i^{\text{cal}})\}_{i=1}^n$. The reference sample $\mathcal{D}_{\text{ref}}$ may be used to construct query-dependent uncertainty sets or scores and is fixed before conformal calibration. The calibration observations and a future test observation $(X_{\text{new}}, Y_{\text{new}})$ are assumed exchangeable conditional on everything learned from $\mathcal{D}_{\text{ref}}$. For each query x, we wish to construct a measurable nested family of residual sets $\widehat{G}_a(x) \subseteq \mathcal{R}$, indexed by $a \geq 0$, i.e., $\widehat{G}_b(x) \subseteq \widehat{G}_a(x)$ for all $0 \leq b \leq a$. Here, $\widehat{G}_a(x)$ represents the residuals regarded as plausible at inflation level a. This family induces the scalar nonconformity score $\widehat{S}(x, y) := \inf\{a \geq 0 : T(x, y) \in \widehat{G}_a(x)\}$. Thus, $\widehat{S}(x, y)$ is the smallest inflation level for which the residual of (x, y) enters the query-dependent set. Given a target miscoverage level $\alpha \in (0, 1)$, the calibration scores $\widehat{S}_i = \widehat{S}(X_i^{\text{cal}}, Y_i^{\text{cal}})$ yield the split-conformal threshold $\widehat{q}_\alpha$ and resulting query-dependent prediction set $\widehat{C}_\alpha(x) := \{y \in \mathcal{Y} : \widehat{S}(x, y) \leq \widehat{q}_\alpha\} = \{y \in \mathcal{Y} : T(x, y) \in \widehat{G}_{\widehat{q}_\alpha}(x)\}$.

The primary requirement we desire is *finite-sample marginal validity*: $\mathbb{P}(Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}})) \geq 1 - \alpha$. Beyond marginal validity, we seek strong *local* behavior. Define the conditional score CDF $F_x(a) := \mathbb{P}(\widehat{S}(X, Y) \leq a \mid X = x) = P_x^R(\widehat{G}_a(x))$. Ideally, $F_x(a)$ is *invariant* across x: for each inflation level a, $\widehat{G}_a(x)$ contains the same conditional residual probability mass at every query. More generally, we seek $F_x(a) \approx F_{x'}(a)$ for relevant x, x', a, equivalently $P_x^R(\widehat{G}_a(x)) \approx P_{x'}^R(\widehat{G}_a(x'))$. Under this score comparability, conditioned on the fitted score and calibration sample, local coverage at query x is $F_x(\widehat{q}_\alpha)$; when the conditional score laws are comparable, the same fitted threshold has similar probability content across queries and is approximately $1 - \alpha$.

The goal is to construct query-adaptive residual geometries that preserve finite-sample marginal validity while making the induced scalar scores sufficiently comparable for globally pooled calibration to remain locally meaningful. We seek: (i) *adaptive geometry*, so $\widehat{G}_a(x)$ captures the conditional structure of P_x^R beyond a scalar scale; (ii) *score comparability*, so nonconformity distributions vary little across queries; (iii) *calibration efficiency*, avoiding separate quantile estimates from small local subsets; and (iv) *scalability*, so adaptation depends on uncertainty-relevant variation rather than ambient dimension. The resulting sets should admit global conformal calibration while remaining marginally valid and, with accurate geometry, approximately conditionally valid.

4 METHOD: FACCP

We overview FacCP (see Fig. 1), which separates two statistical tasks. Auxiliary data determine how residual behavior changes with the query, while an independent calibration sample determines only a single global inflation level. We describe how the conditional residual law is estimated (Sec. 4.1), how prediction sets are constructed from the retrieved law (Sec. 4.2), and show when a single global conformal threshold is locally meaningful (Sec. 4.3).

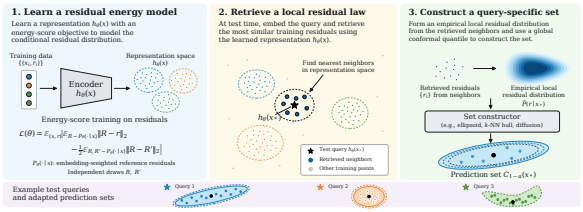


Figure 1: Overview of FacCP.

4.1 LEARNING THE CONDITIONAL RESIDUAL LAW

The first step of FacCP is to estimate how prediction errors vary with the query, i.e., the conditional residual law P_x^R in (1). When P_x^R varies with x, a fixed residual geometry can be inefficient because it must accommodate errors that are plausible for some queries but unlikely for others. FacCP therefore first estimates P_x^R as $\widehat{P}_x^R$ and then maps $\widehat{P}_x^R$ to a query-dependent residual set $\widehat{G}_a(x)$. This separates the statistical problem of learning query-dependent uncertainty from the downstream choice of geometric representation, enabling the same estimated residual law to support different set families such as ellipsoids or non-convex sets without retraining.

Residual law retrieval. Using $\mathcal{D}_{\text{ref}}$, where $R_j = T(X_j^{\text{ref}}, Y_j^{\text{ref}})$, we estimate P_x^R as an input-dependent reweighting of observed residuals from $\mathcal{D}_{\text{ref}}$. Direct input-space distances can be misleading when geometrically dissimilar inputs induce similar predictive uncertainty, e.g., rotated images with unchanged class uncertainty. We thus learn a representation $h_\theta : \mathcal{X} \to \mathbb{R}^p$, writing $Z = h_\theta(X)$ and $z = h_\theta(x)$, which maps queries that induce similar conditional residual laws to be close in the learned representation. Given a query x, define normalized retrieval weights

$$w_j(x) = \frac{\exp\{-d_Z(h_\theta(x), h_\theta(X_j))/\tau\}}{\sum_{k=1}^m \exp\{-d_Z(h_\theta(x), h_\theta(X_k))/\tau\}}, \qquad \sum_{j=1}^m w_j(x) = 1, \tag{2}$$

where d_Z is a metric in latent space and $\tau > 0$ controls the locality of the weighting. We use the weighted empirical residual distribution as our resulting estimator:

$$\widehat{P}_x^R = \sum_{j=1}^m w_j(x)\delta_{R_j}. \tag{3}$$

Note that the residuals themselves are not compressed into the latent space, and $\widehat{P}_x^R$ remains a distribution on the original residual space $\mathcal{R}$; h_θ is used only to determine the residual weights when estimating the conditional distribution P_x^R. Thus, the role of the latent space is to solve a localization problem, identifying reference queries whose residual distributions are informative about P_x^R.

Learning an uncertainty-relevant metric. It remains to determine how h_θ should be trained. A standard representation loss based on input-space proximity is insufficient because nearby inputs can have different residual laws, while distant inputs can exhibit similar uncertainty. We therefore train the representation according to the quality of the residual distribution it retrieves. For a training sample (x_i, r_i), let J_i denote reference indices excluding the query itself. Define weights $w_{ij}^{(\theta)}$ and the corresponding retrieved distribution as $\widehat{P}_{x_i,\theta}^R$:

$$w_{ij}^{(\theta)} = \frac{\exp\{-d_Z(h_\theta(x_i), h_\theta(X_j))/\tau\}}{\sum_{k\in J_i} \exp\{-d_Z(h_\theta(x_i), h_\theta(X_k))/\tau\}}, \qquad j \in J_i, \qquad \widehat{P}_{x_i,\theta}^R = \sum_{j\in J_i} w_{ij}^{(\theta)}\delta_{R_j}, \tag{4}$$

The exclusion of observation i prevents the representation from obtaining a trivial solution by retrieving the residual r_i associated with the query itself. We train h_θ by minimizing

$$\ell_i(\theta) = \sum_{j\in J_i} w_{ij}^{(\theta)} \|R_j - r_i\|_2 - \tfrac{1}{2}\sum_{j,k\in J_i} w_{ij}^{(\theta)} w_{ik}^{(\theta)} \|R_j - R_k\|_2. \tag{5}$$

The first term favors retrieved residuals compatible with r_i, but alone would collapse the retrieved law toward the nearest reference residual. The second term prevents this by rewarding the dispersion needed to represent a nondegenerate distribution. Thus, (5) scores the retrieved residuals collectively as a probability law. No residual covariance or scale is given during training; each r_i is simply one draw from the unknown conditional law $P_{x_i}^R$, which the retrieved distribution is trained to predict.

Energy-score interpretation. We justify (5) as a principled training loss for estimating conditional residual laws. For a training query x_i, the representation induces the weighted empirical law $\widehat{P}_{x_i,\theta}^R$ (4). We show that (5) is exactly the *energy score* of this retrieved law evaluated at the observed residual r_i. For a law Q on $\mathcal{R} \subseteq \mathbb{R}^{d_R}$ with finite first moment and an observation $r \in \mathcal{R}$, define the energy score $\text{ES}(Q, r) = \mathbb{E}_{V\sim Q}\|V - r\|_2 - \frac{1}{2}\mathbb{E}_{V,V'\overset{\text{i.i.d.}}{\sim} Q}\|V - V'\|_2$. The first term rewards distributions placing mass near the realized residual, while the second prevents a degenerate solution that concentrates all probability mass at a single point. For the retrieved residual law $\widehat{P}_{x_i,\theta}^R$, direct substitution into the energy-score definition gives $\ell_i(\theta) = \text{ES}(\widehat{P}_{x_i,\theta}^R, r_i)$; the formal statement and proof are given in Prop. 5. As the energy score is strictly proper (Gneiting & Raftery, 2007, Thm. 5), its expected value is uniquely minimized by the true data-generating law: $P = \arg\min_Q \mathbb{E}_{R\sim P}[\text{ES}(Q, R)]$. Applied conditionally, the unrestricted population target of (5) is thus P_x^R almost surely under P_X. Thus, (5) targets the full conditional residual law (not just selected moments); the retrieval architecture only restricts which such laws can be represented in practice.

Uncertainty sufficiency. If residual retrieval uses only $Z = h_\theta(X)$, what information about X must Z preserve? For the purpose of modeling the conditional residual distribution, Z need not retain all information needed to predict Y; rather, it only needs to preserve the information determining the conditional residual law. We call Z *uncertainty-sufficient* if $R \perp\!\!\!\perp X \mid Z$, or equivalently, since Z is a deterministic function of X, if $\mathcal{L}(R \mid X) = \mathcal{L}(R \mid Z)$ almost surely. Thus, once Z is known, the remaining information in X does not further change the distribution of R.

In practice, $h_\theta(X)$ may not be exactly uncertainty-sufficient, so to quantify departures from this ideal, we define $P_X^R := \mathcal{L}(R \mid X)$ and $P_Z^R := \mathcal{L}(R \mid Z)$. Here, P_X^R is the residual law conditioned

on the full query, whereas P_Z^R is the law available after compressing X to Z. If compression discards uncertainty-relevant information, these laws differ. For probability laws P, Q with finite first moments, the energy score satisfies the standard identity $\mathbb{E}_{R\sim P}[\mathrm{ES}(Q,R)] - \mathbb{E}_{R\sim P}[\mathrm{ES}(P,R)] = \frac{1}{2}\mathcal{E}(P,Q)$, where $\mathcal{E}(P,Q) \geq 0$ is the energy distance and equals zero if and only if $P = Q$. Applying this identity to P_X^R and P_Z^R gives the following representation-specific result:

Proposition 1 (Energy-score cost)**.** *Assume the conditional residual laws have finite first moments, and define* $\mathcal{L}_X^\star = \mathbb{E}[\mathrm{ES}(P_X^R, R)]$ *and* $\mathcal{L}_Z^\star = \mathbb{E}[\mathrm{ES}(P_Z^R, R)]$*. Then* $\mathcal{L}_Z^\star - \mathcal{L}_X^\star = \frac{1}{2}\mathbb{E}\big[\mathcal{E}(P_X^R, P_Z^R)\big]$*. Thus* $\mathcal{L}_Z^\star = \mathcal{L}_X^\star$ *if and only if* $P_X^R = P_Z^R$ *almost surely, equivalently, if and only if* $R \perp\!\!\!\perp X \mid Z$*.*

Prop. 1 shows compression is lossless when Z is uncertainty-sufficient; otherwise, it incurs excess energy-score risk equal to $\frac{1}{2}\mathbb{E}\big[\mathcal{E}(P_X^R, P_Z^R)\big]$. Thus, h_θ should preserve variations in X that change the residual law while discarding those that do not. If these factors are low-dimensional, a compact Z can preserve residual-law information while making localization statistically easier.

Intrinsic dimension and scalability. Uncertainty sufficiency explains when localization in a compressed representation preserves the conditional residual law. Compression can also improve the statistical efficiency of residual-law estimation by removing directions that affect retrieval distances but not the conditional residual law. Under a stylized local-kernel model with regular local geometry and Hölder-smooth conditional residual laws, formalized in Prop. 6, suppose an uncertainty-sufficient localization variable has intrinsic dimension d, while $p - d$ additional directions vary under the retrieval metric but are conditionally irrelevant to R. Writing $\mathfrak{R}_s^\star$ for the optimally tuned worst-case expected energy-distance error when localizing in intrinsic dimension s, the proposition gives $\mathfrak{R}_d^\star \asymp m^{-2\beta/(2\beta+d)}$ and $\mathfrak{R}_p^\star \asymp m^{-2\beta/(2\beta+p)}$ for $p > d$. Thus, uncertainty-irrelevant variation slows residual-law estimation despite carrying no additional information about the residual law (we validate this empirically in Fig. 2). Learned compression can therefore improve residual-law estimation by suppressing nuisance variation, but excessive compression can discard information that changes P_x^R, incurring the representation error characterized in Prop. 1.

4.2 From Residual Distributions to Conformal Sets

The retrieved conditional law $\widehat{P}_x^R$ describes plausible residuals at query x, but is not itself a prediction set. We use it to construct a *query-dependent geometry* and use conformal calibration to determine the inflation required for the desired coverage. This separates *local adaptation* from *calibration*, allowing the geometry to vary with x while retaining a single scalar nonconformity score calibrated using the full calibration set $\mathcal{D}_{\mathrm{cal}}$. Formally, let Γ be a *geometry constructor* that maps any probability law Q on $\mathcal{R}$ to a nested family $\{\Gamma_a(Q)\}_{a\geq 0}$ satisfying $\Gamma_b(Q) \subseteq \Gamma_a(Q)$ for all $0 \leq b \leq a$. Given the retrieved residual law $\widehat{P}_x^R$, we define the local residual geometry as $\widehat{G}_a(x) := \Gamma_a(\widehat{P}_x^R)$. Nestedness ensures that increasing a only enlarges the admissible residual set, so the family can be represented by a scalar nonconformity score. The form of Γ can otherwise be chosen to match the residual space and downstream application. For example, an ellipsoidal constructor summarizes $\widehat{P}_x^R = \sum_{j=1}^m w_j(x)\delta_{R_j}$ through its weighted center $\widehat{c}(x) = \sum_{j=1}^m w_j(x)R_j$ and regularized weighted covariance $\widehat{\Sigma}(x) = \sum_{j=1}^m w_j(x)\big(R_j - \widehat{c}(x)\big)\big(R_j - \widehat{c}(x)\big)^\top + \lambda I$, for $\lambda > 0$. The resulting nested family is $\widehat{G}_a(x) = \{r \in \mathcal{R} : (r - \widehat{c}(x))^\top \widehat{\Sigma}(x)^{-1}(r - \widehat{c}(x)) \leq a^2\}$. Thus $\widehat{P}_x^R$ determines the local center, orientation, and relative scale, while a controls only overall inflation. The same retrieved law can support other nested families, such as hyperboxes or zonotopes, without retraining h_θ, since the representation targets the residual law rather than a specific geometry. We then map this query-dependent geometry to a scalar nonconformity score, defined for candidate response y as the smallest inflation level containing its residual:

$$\widehat{S}(x,y) := \inf\{a \geq 0 : T(x,y) \in \widehat{G}_a(x)\}. \tag{6}$$

Although the geometry used to measure the residual can vary substantially with x, every observation is now represented by a scalar score. For the constructors considered here, the infimum is attained and the nested family is right-continuous, ensuring that $\widehat{S}(x,y) \leq a$ is equivalent to $T(x,y) \in \widehat{G}_a(x)$.

This equivalence lets conformal calibration of scalar scores map directly back to residual sets. We then freeze $\widehat{S}$ and use an independent calibration sample $\mathcal{D}_{\mathrm{cal}} = (X_i, Y_i)_{i=1}^n$. Since query dependence is already encoded in $\widehat{S}$, no further local fitting or dataset partitioning is needed, and all n calibration points contribute to the same conformal quantile despite having different local

geometries. Compute $S_i = \widehat{S}(X_i, Y_i)$ for every calibration observation, let $S_{(1)} \le \cdots \le S_{(n)}$ denote the ordered scores, and define $S_{(n+1)} := \infty$. For target miscoverage $\alpha \in (0,1)$, let $k = \lceil (n+1)(1-\alpha) \rceil$ and set $\widehat{q}_\alpha := S_{(k)}$. The resulting conformal prediction set is $\widehat{C}_\alpha(x) = \{y \in \mathcal{Y} : \widehat{S}(x,y) \le \widehat{q}_\alpha\} = \{y \in \mathcal{Y} : T(x,y) \in \widehat{G}_{\widehat{q}_\alpha}(x)\}$. This construction separates adaptation from calibration: the retrieved residual law determines the query-dependent geometry $\widehat{G}_a(x)$, while all calibration points estimate a single global inflation level $\widehat{q}_\alpha$. Thus, FacCP preserves the full calibration sample rather than estimating local quantiles. Once $\widehat{S}$ is fixed using data independent of $\mathcal{D}_{\mathrm{cal}}$, standard split-conformal validity gives, under exchangeability, $\mathbb{P}(Y_{\mathrm{new}} \in \widehat{C}_\alpha(X_{\mathrm{new}})) \ge 1-\alpha$; a formal conditional statement is given in Thm. 7. This validity does not depend on the accuracy of $\widehat{P}_x^R$ or $\widehat{G}_a(x)$: retrieval quality instead determines how well the geometry adapts across queries and hence the efficiency and comparability of the resulting scores.

4.3 Pivotality and Local Coverage

The previous section showed that a single global conformal quantile $\widehat{q}_\alpha$ guarantees finite-sample marginal coverage regardless of geometry quality. Here we seek the stronger property that the same threshold has approximately the same probabilistic meaning across queries, allowing calibration scores to be pooled without sacrificing local adaptivity. The relevant property is *pivotality*. Throughout this subsection, we condition on $\mathcal{D}_0$, containing all data and randomness used to construct $\widehat{S}$ in (6), and suppress this conditioning; hence $\widehat{S}$ is fixed. Define the conditional and marginal score distributions by $F_x(a) := \mathbb{P}(\widehat{S}(X,Y) \le a \mid X = x)$ and $F(a) := \mathbb{P}(\widehat{S}(X,Y) \le a)$. Let $\mathcal{X}_0 \subseteq \mathcal{X}$ denote the covariate region over which local behavior is assessed; we take $\mathcal{X}_0$ to contain the support of the marginal distribution defining F, as well as any target-relevant covariates considered under shift.

Definition 2 (Pivotality)**.** The score $\widehat{S}(X,Y)$ is pivotal with respect to X on $\mathcal{X}_0$ if $F_x(a) = F(a)$ for every $x \in \mathcal{X}_0$ and $a \ge 0$. We quantify departures from pivotality by $\Delta_{\mathrm{piv}} := \sup_{x \in \mathcal{X}_0,\, a \ge 0} |F_x(a) - F(a)|$.

Pivotality does not require identical set shapes across queries; it requires each inflation level a to contain approximately the same conditional residual mass. If $\Delta_{\mathrm{piv}} = 0$, one global threshold has the same probability content everywhere. Thus, FacCP absorbs query-dependent variation into the learned geometry before scoring, making scores comparable across queries and allowing the full calibration sample to share a single global quantile. With the covariate region $\mathcal{X}_0$ defined above, consider an oracle nested geometry $\{G_a^\star(x)\}_{a \ge 0}$ with query-invariant conditional mass $P_x^R(G_a^\star(x)) = F_0(a)$, for all $x \in \mathcal{X}_0,\ a \ge 0$. Such an oracle can be obtained by mapping each P_x^R to a common reference law and applying a common nested family in those coordinates; the formal construction and oracle-pivotality statement are given in Thm. 9. FacCP does not estimate this coordinate map explicitly: the retrieved residual law instead determines a query-dependent geometry directly. Perfect recovery of the oracle geometry would yield exact pivotality; in practice, it suffices to approximate the oracle where P_x^R places probability mass.

Assumption 3 (Uniform geometry approximation)**.** *For some* $\varepsilon_{\mathrm{geom}} \ge 0$, $\sup_{x \in \mathcal{X}_0,\, a \ge 0} P_x^R\Big(\widehat{G}_a(x) \triangle G_a^\star(x)\Big) \le \varepsilon_{\mathrm{geom}}$*, where* $\triangle$ *denotes symmetric difference.*

The quantity $\varepsilon_{\mathrm{geom}}$ is the worst-case conditional probability mass on which the learned and oracle sets disagree, ignoring discrepancies in negligible-probability regions. It links representation learning to coverage: better residual-law retrieval yields smaller $\varepsilon_{\mathrm{geom}}$ when Γ is stable to errors in $\widehat{P}_x^R$.

Theorem 4 (Geometry approximation controls pivotality)**.** *Under Assumption 3,* $\sup_{x \in \mathcal{X}_0,\, a \ge 0} |F_x(a) - F_0(a)| \le \varepsilon_{\mathrm{geom}}$*, and* $\Delta_{\mathrm{piv}} \le 2\varepsilon_{\mathrm{geom}}$.

Thm. 4 clarifies the role of training: the loss (5) targets P_x^R, Γ converts it into an adaptive residual set, and under Assumption 3, local geometry error directly bounds score non-pivotality. Pivotality then makes scores comparable across queries, allowing the *full* calibration set to estimate one threshold that remains approximately valid locally.

Define the marginal population quantile $q_\alpha := F^{-1}(1-\alpha) := \inf\{a : F(a) \ge 1-\alpha\}$, where F^{-1} is the generalized inverse, and suppose $F(q_\alpha) = 1-\alpha$. Then

$$\sup_{x \in \mathcal{X}_0} |\mathbb{P}(\widehat{S}(X,Y) \le q_\alpha \mid X = x) - (1-\alpha)| \le \Delta_{\mathrm{piv}}. \tag{7}$$

See Cor. 10 for a formal proof. Thus Δ_{piv} is the structural worst-case local coverage error remaining even with the exact marginal population quantile. Finite calibration adds only the error from estimating this global quantile, of order $n^{-1/2}$; a formal high-probability statement is given in Thm. 11. Combined with Thm. 4, this separates local error into geometry-induced non-pivotality and global calibration error. As Δ_{piv} is a population supremum rather than a directly estimable uniform certificate in general continuous settings, we use it primarily as an empirical diagnostic.

Transfer under covariate shift. Pointwise conditional coverage can equivalently be viewed as marginal coverage under covariate reweightings that preserve $P_{Y|X}$ (Gibbs et al., 2025). Accordingly, approximate pivotality also supports transfer of the same unweighted global threshold across covariate shifts that preserve the conditional response law. A formal finite-calibration result and comparison with shift-aware calibration are given in Cor. 12.

5 Numerical Experiments

5.1 Controlled Residual Geometry and Localization

We first isolate the role of residual geometry and localization in a setting with an analytic conditional residual law. Let $X = (X_1, X_2, X_3, N)$, where the first three coordinates independently control residual scale, aspect ratio, and orientation, and N contains optional nuisance coordinates independent of the residual law. We draw $R \mid X = x \sim \mathcal{N}(0, \Sigma(x))$. The nuisance coordinates affect distances in the observed input space but do not affect $\Sigma(x)$ or the conditional residual law P_x^R.

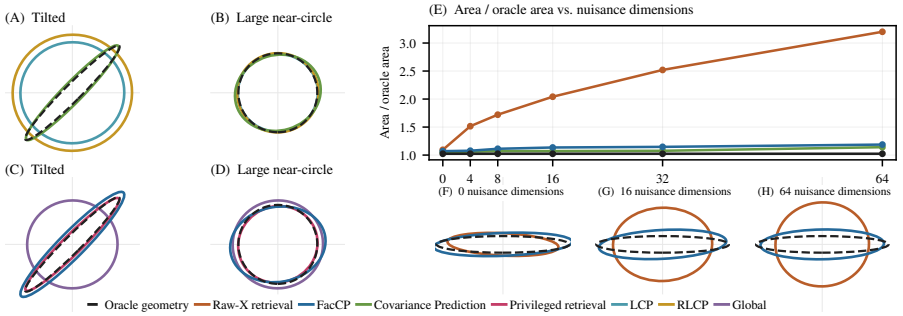


Figure 2: Controlled residual-geometry experiment. (A–D) show prediction sets under multiple methods. (E) shows that FacCP stays close to oracle area compared to raw-X retrieval under nuisance dimensions. (F–H) show the degradation of raw-X retrieval at multiple nuisance dimensions.

Fig. 2 compares six geometry constructions. Oracle uses the true $\Sigma(x)$, Global uses one covariance for all queries, Privileged retrieval localizes with the known uncertainty-generating coordinates, Raw-X retrieval localizes directly in X, FacCP learns the retrieval representation via the energy-score objective, and covariance prediction directly estimates $\Sigma(x)$ by Gaussian likelihood. Without nuisance coordinates, FacCP achieves mean area $1.07\times$ oracle, compared with $3.95\times$ for Global. Panels A–B show that quantile localization (LCP (Guan, 2023), RLCP (Hore & Barber, 2025)) can resize an ellipse, but cannot recover changes in aspect ratio or orientation. Score-pivotality diagnostics and baseline results are reported in the Appendix C.1. Raw-X retrieval works well without nuisance coordinates but degrades as uncertainty-irrelevant variation grows. At 16 nuisance dimensions, its mean area is $2.04\times$ oracle with worst conditional-coverage deviation 0.301, versus $1.14\times$ and 0.066 for FacCP; at 64 dimensions, raw-X reaches $3.20\times$ while FacCP remains at $1.19\times$. Panels E–H illustrate the corresponding area and geometry degradation, consistent with Prop. 6. Covariance prediction remains strong in this data-rich, well-specified setting; Sec. 5.2 studies lower-data and lower-capacity regimes. Localized-CP methods reduce heterogeneity in the global score but remain limited to globally oriented ellipses. Even with the pointwise conditional 95% score quantile, mean area is $1.99\times$ oracle, showing that local calibration corrects scalar variation but not query-dependent aspect ratio or orientation.

Covariate Shift We test the covariate-shift transfer result of Cor. 12 by calibrating once on the source distribution and evaluating under shifts that preserve $P_{R|X}$. With $U_1, U_2, U_3 \sim \text{Unif}[-1, 1]$ controlling residual scale, aspect ratio, and orientation, we tilt one coordinate at test time via $p_\beta(u_j) \propto \exp(\beta u_j)$, with $\beta = 0$ recovering the source. Under scale and aspect-ratio shifts, Global coverage degrades to 87.3% and 89.2% at $\beta = 3$, while FacCP remains near the nominal 95% target at 94.9% and 96.4%. Under orientation shift, both remain near nominal coverage because the pooled Global geometry is roughly rotation-symmetric here. Full sweeps and baselines are in App. C.2.

5.2 QUADCOPTER REACHABILITY

2D Quadcopter. We evaluate FacCP on 100-step planar quadcopter rollouts. Each query contains a nominal 6-state trajectory and a 2-dimensional control sequence, yielding an 806-dimensional query, while the residual trajectory contains 600 coordinates. We target simultaneous 99% coverage of the full trajectory. Rather than calibrating each time step separately and applying a Bonferroni correction, we follow the single nonconformity score approach of Lindemann et al. (2025), which aggregates time-indexed geometric scores into one trajectory-level score and conformalizes a single threshold. FacCP changes the geometry entering this score. The global construction uses one residual geometry for all queries, while FacCP constructs the geometry from the retrieved conditional residual law and applies the same trajectory-level conformal calibration. This gives a high-dimensional local conformal prediction problem without splitting calibration support across the 806-dimensional query space (Guan, 2023; Hore & Barber, 2025). See App. C.3 for details.

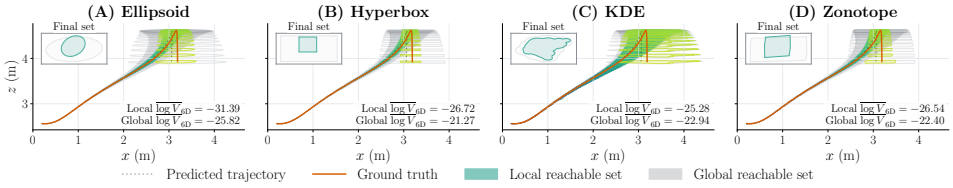


Figure 3: One learned representation supports multiple geometries without retraining.

Fig. 3 shows that one h_θ supports ellipsoidal, hyperbox, KDE, and zonotopic sets without retraining. Across constructors, local geometry reduces log volume by 2.34 to 5.57 relative to global geometry. For ellipsoids, retrieval in the raw and standardized 806D input gives log volumes -26.66 and -26.84, close to the global -25.82, whereas learned retrieval reaches -33.25 with 99.1% coverage. This matches Prop. 1 and Prop. 6. An uncertainty-sufficient representation discards variation that does not change the residual law, while removing such variation from the retrieval metric increases effective local support and lowers the intrinsic localization dimension. Thus FacCP solves the same residual-law estimation problem in a more favorable localization space. See App. C.3 for details.

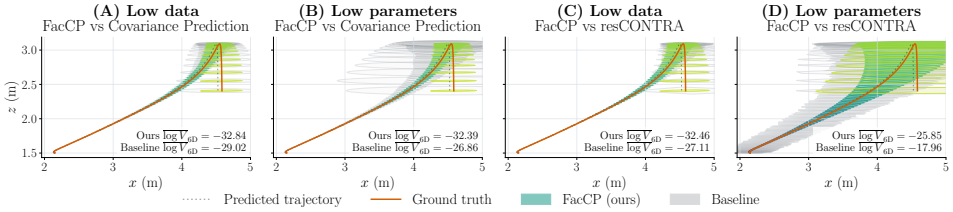


Figure 4: FacCP outperforms baselines in low data and model parameter regimes.

We compare FacCP with resCONTRA (Fang et al., 2025) and Covariance Prediction CP (CVCP) (Srinivasan et al., 2026), matching learned parameter counts across capacity-sweep points; full results appear in Table 7. resCONTRA learns a query-dependent flow, while CVCP directly predicts covariance. At roughly 5×10^3 parameters, FacCP yields sets over $10^3\times$ smaller than resCONTRA at comparable coverage, and at low capacity over $250\times$ smaller than CVCP (Figure 4). These gaps close only with substantially greater direct-model capacity. Prop. 5 shows that the retrieval objective is a proper scoring rule whose unrestricted population target is the conditional residual law; this encourages pooling reference residuals with similar conditional laws rather than directly encoding the final covariance or flow. We next fix capacity near each sweep's crossover and vary the training-data budget; full results are in Table 8. At 20% of the data, FacCP produces sets $211\times$ smaller than resCONTRA and $46\times$ smaller than CVCP while maintaining coverage (Figure 4); these gaps shrink with more data. Consistent with Props. 5, 1, and 6, FacCP uses capacity to learn an uncertainty-relevant localization metric and constructs geometry from retrieved residuals, whereas direct methods parameterize the query-dependent geometry itself. This reduces FacCP's estimation burden in low-data, low-capacity regimes, with the advantage shrinking as direct estimators gain sufficient data and capacity.

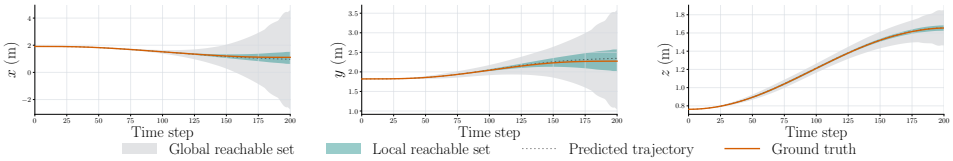


Figure 5: 3D quadcopter reachability over 200 steps.

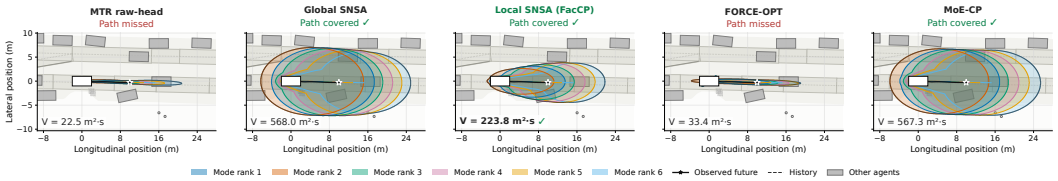


Figure 6: A scene with at least seven neighboring agents at the 95% target. Local SNSA (FacCP) achieves lowest volume while achieving coverage.

3D Quadcopter. We scale the planar experiment to 200-step rollouts of a 12-state, 4-control 3D quadcopter, yielding a 3212-dimensional query and a 2400-dimensional residual trajectory. The construction and trajectory-level calibration are unchanged from the planar setup above, with a single conformal event targeting 99% simultaneous coverage of the full trajectory. Figure 5 shows that local retrieval remains effective at this scale: mean per-time-step 12D log volume drops from -52.52 to -68.99, with essentially unchanged coverage (99.4% global vs. 99.3% local).

The capacity and data sweeps sharpen this scaling result. FacCP reaches log volumes near -78 with tens of thousands of parameters; Mahalanobis CP remains less efficient even at $\sim 1.6 \times 10^6$ parameters and across every tested data budget. resCONTRA closes the capacity gap between 6×10^4 and 9×10^4 parameters and, at fixed $\sim 6 \times 10^4$ capacity, closes the data gap near 70% of the training budget. Thus retrieval is more efficient in the low-capacity and low-data regimes, while an expressive and sufficiently trained flow can eventually surpass it. Full sweeps are in App. C.3.1.

5.3 Autonomous Driving

We evaluate FacCP around a frozen MTR forecaster trained with annealed Winner-Takes-All (Shi et al., 2022; Xu et al., 2025) on Argoverse 2 (Wilson et al., 2021). Each scene contains agent histories and a local road map. MTR predicts six candidate trajectories over six seconds, each with a probability. We keep these paths fixed and construct uncertainty tubes around them. Our guarantees require one of the six tubes to contain the complete observed future at all times without switching between modes. A single-mode comparison, using only the most-probable path, is reported in Appendix D.6.

Method	Footprint (m^2·s) ± SD	Whole-path coverage (%)
MTR raw-head	1,056.0	4.15
Global SNSA	683.8 ± 9.6	95.19
Local SNSA (FacCP)	581.9 ± 11.3	94.46
FORCE-OPT	197.6	62.29
MoE-CP	766.1 ± 10.8	95.29

Table 1: Six-mode results at 95%.

FORCE-OPT constructs compact reachable sets for motion-plan safety evaluation (Chakraborty et al., 2026), but calibrates per-timestep coverage. Our SNSA-based trajectory score instead calibrates coverage by one complete candidate tube (App. D.3). MoE-CP adapts calibration through expert-gating similarity (Kong et al., 2026); FacCP instead adapts tube geometry through learned scene–candidate similarity, followed by ordinary split-conformal calibration. Global SNSA estimates timestep-wise covariance and residual scales using equal reference weights; Local SNSA (FacCP) uses learned embedding weights. For each candidate, we take the maximum normalized error over time and add the pricing penalty, then minimize over candidates and calibrate. Localization adapts geometry to scene-dependent errors and pricing allocates radius budgets by candidate probability. Implementation details are in App. D.2 and D.4; App. D.8 isolates pricing. We report footprint—time-integrated spatial area counting modal overlap once—and whole-path coverage. Refer appendix D.1). At 95%, priced Local SNSA reduces mean footprint by 14.9% relative to priced Global SNSA (Table 1). FORCE-OPT' yields smaller sets (197.6), but only 62.29% single-tube whole-path coverage despite its per-timestep calibration. Fig. 6 illustrates smaller local footprints covering complete trajectories in crowded ($\geq$ 7 neighbors) scenes. Full results are in App. D.8, D.5, and D.6.

6 Conclusion

FacCP reframes adaptive conformal prediction around estimating the conditional residual law, rather than directly learning a score, quantile, or final set. This factorization lets query-dependent geometry adapt locally while retaining a single global conformal calibration, and allows one learned representation to support multiple set families. Across high-dimensional reachability and autonomous driving, FacCP produces tighter calibrated sets, with particularly strong gains in data- and parameter-limited regimes.

Limitations. Our local-coverage guarantees depend on the pivotality defect, which is a population quantity that we currently assess empirically rather than certify with finite-sample bounds. Developing practical upper bounds on this defect—and thereby turning empirical pivotality diagnostics into quantitative local-coverage guarantees—is an important direction for future work.

AI USE STATEMENT

Generative AI tools were used in the preparation of this work to assist with refining mathematical statements and annotations, providing feedback on human-designed experimental methodology and ablation, and improving the organization, clarity, and presentation of the manuscript. They were also used for language editing, restructuring sections, and formatting equations and tables.

All AI-assisted mathematical arguments were reviewed and revised by the authors, and all reported experimental results were produced and verified by the authors. Generative AI was not treated as a source of scientific evidence or authority; technical claims, derivations, and conclusions were checked against the underlying analysis and experimental results. The authors take full responsibility for the final content of the paper, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Salim I Amoukou and Nicolas JB Brunel. Adaptive conformal prediction by reweighting nonconformity score. *arXiv preprint arXiv:2303.12695*, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi: 10.1093/imaiai/iaaa017.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction, 2025. URL https://arxiv.org/abs/2512.11779.

Sacha Braun, Liviu Aolaritei, Michael I Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *Journal of the American Statistical Association*, (just-accepted): 1–27, 2026.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021. URL https://www.jmlr.org/papers/v22/20-753.html.

Kaustav Chakraborty, Zeyuan Feng, Sushant Veer, Apoorva Sharma, Wenhao Ding, Sever Topan, Boris Ivanovic, Marco Pavone, and Somil Bansal. Safety evaluation of motion plans using trajectory predictors as forward reachable set estimators. *IEEE Robotics and Automation Letters*, 2026.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021. doi: 10.1073/pnas.2107794118.

Nicolò Colombo. Normalizing flows for conformal regression. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 881–893. PMLR, 2024. URL https://proceedings.mlr.press/v244/colombo24a.html.

Nicolas Deutschmann, Mattia Rigotti, and Maria Rodriguez Martinez. Adaptive conformal regression with split-jackknife+ scores. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=1fbTGC3BUD.

Victor Dheur, Tanguy Bosser, Rafael Izbicki, and Souhaib Ben Taieb. Distribution-free conformal joint prediction regions for neural marked temporal point processes. *Machine Learning*, 113(9): 7055–7102, 2024. doi: 10.1007/s10994-024-06594-z.

Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A unified comparative study with generalized conformity scores for multi-output conformal regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume

267 of *Proceedings of Machine Learning Research*, pp. 13444–13485. PMLR, 2025. URL https://proceedings.mlr.press/v267/dheur25a.html.

Nathaniel Diamant, Ehsan Hajiramezanali, Tommaso Biancalani, and Gabriele Scalia. Conformalized deep splines for optimal and efficient prediction sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1657–1665. PMLR, 2024. URL https://proceedings.mlr.press/v238/diamant24a.html.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64555–64576, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/cb931eddd563f8d473c355518ce8601c-Abstract-Conference.html.

Eshant English and Christoph Lippert. Japan: Joint adaptive prediction areas with normalising flow. In *International Conference on Learning Representations*, volume 2026, pp. 131010–131047, 2026.

Zhenhan Fang, Aixin Tan, and Jian Huang. Contra: Conformal prediction region via normalizing flow transformation. In *International Conference on Learning Representations*, volume 2025, pp. 91833–91850, 2025.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023. URL https://www.jmlr.org/papers/v24/21-1280.html.

Ruijiang Gao, Mingzhang Yin, James McInerney, and Nathan Kallus. Adjusting regression models for conditional uncertainty calibration. *Machine Learning*, 113(11):8347–8370, 2024. doi: 10.1007/s10994-024-06627-7.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction, 2022. URL https://arxiv.org/abs/2206.13092.

Jonas Hanselle, Alireza Javanmardi, Tobias Florin Oberkofler, Yusuf Sale, and Eyke Hüllermeier. Conformal prediction without nonconformity scores. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1626–1639. PMLR, 2025. URL https://proceedings.mlr.press/v286/hanselle25a.html.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2025.

Rafael Izbicki, Gilson Shimizu, and Rafael Stern. Flexible distribution-free conditional predictive bands using density estimators. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3068–3077. PMLR, 2020. URL https://proceedings.mlr.press/v108/izbicki20a.html.

Rafael Izbicki, Gilson Shimizu, and Rafael B. Stern. CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022. URL https://www.jmlr.org/papers/v23/20-797.html.

Sunay Joshi, Shayan Kiyani, George J Pappas, Edgar Dobriban, and Hamed Hassani. Conformal inference under high-dimensional covariate shifts via likelihood-ratio regularization. *Advances in Neural Information Processing Systems*, 38:16079–16120, 2026.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24749–24769. PMLR, 2024. URL https://proceedings.mlr.press/v235/kiyani24a.html.

Jingsen Kong, Wenlu Tang, Dezheng Kong, Linglong Kong, Guangren Yang, and Bei Jiang. Adaptive conformal prediction via mixture-of-experts gating similarity. In *International Conference on Learning Representations*, volume 2026, pp. 94169–94191, 2026.

Félix Laplante. A post-processing conformal prediction approach for conditional coverage via pivotal scores, 2026. URL https://arxiv.org/abs/2605.25852.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014. doi: 10.1111/rssb.12021.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Lars Lindemann, Yiqi Zhao, Xinyi Yu, George J Pappas, and Jyotirmoy V Deshmukh. Formal verification and control with conformal prediction: Practical safety guarantees for autonomous systems. *IEEE Control Systems*, 45(6):72–122, 2025.

Marcos Matabuena, Rahul Ghosal, Pavlo Mozharovskyi, Oscar Hernan Madrid Padilla, and Jukka-Pekka Onnela. Conformal uncertainty quantification using kernel depth measures in separable hilbert spaces, 2024. URL https://arxiv.org/abs/2405.13970.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Conformal and Probabilistic Prediction with Applications*, pp. 294–306. PMLR, 2022.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011. doi: 10.1016/j.neunet.2011.05.008.

Harris Papadopoulos, Alex Gammerman, and Vladimir Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, pp. 64–69, 2008.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49459–49492. PMLR, 2025a. URL https://proceedings.mlr.press/v267/plassier25a.html.

Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9da457059dd3386a2166c5b08f29b7de-Abstract-Conference.html.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 10160–10177. PMLR, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL https://jmlr.org/papers/v9/shafer08a.html.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

Anutam Srinivasan, Antoine Leeman, and Glen Chou. Safety beyond the training data: Robust out-of-distribution mpc via conformalized system level synthesis. *L4DC*, 2026.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59509–59527. PMLR, 2025. URL https://proceedings.mlr.press/v267/thurin25a.html.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Renukanandan Tumu, Matthew Cleaveland, Rahul Mangharam, George Pappas, and Lars Lindemann. Multi-modal conformal prediction regions by optimizing convex shape templates. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1343–1356. PMLR, 2024. URL https://proceedings.mlr.press/v242/tumu24a.html.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490. PMLR, 2012. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8814–8836. PMLR, 2023. URL https://proceedings.mlr.press/v206/wang23n.html.

B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55076–55099. PMLR, 2024. URL https://proceedings.mlr.press/v235/xu24m.html.

Yihong Xu, Victor Letzelter, Mickaël Chen, Éloi Zablocki, and Matthieu Cord. Annealed winner-takes-all for motion forecasting. In *2025 IEEE international conference on robotics and automation (ICRA)*, pp. 1264–1270. IEEE, 2025.

A PROOFS

Proposition 5 (Retrieval loss is the energy score)**.** *For the retrieved residual law* $\widehat{P}^R_{x_i,\theta} = \sum_{j\in J_i} w^{(\theta)}_{ij}\delta_{r_j}$*, the training loss in* (5) *satisfies* $\ell_i(\theta) = \mathrm{ES}(\widehat{P}^R_{x_i,\theta}, r_i)$.

Proof. If $V \sim \widehat{P}^R_{x_i,\theta}$, then $V = r_j$ with probability $w^{(\theta)}_{ij}$. Hence

$$\mathbb{E}\|V - r_i\|_2 = \sum_{j\in J_i} w^{(\theta)}_{ij}\|r_j - r_i\|_2.$$

Likewise, for independent $V, V' \sim \widehat{P}^R_{x_i,\theta}$,

$$\mathbb{E}\|V - V'\|_2 = \sum_{j,k\in J_i} w^{(\theta)}_{ij} w^{(\theta)}_{ik}\|r_j - r_k\|_2.$$

Substituting these expressions into the definition of the energy score yields

$$\mathrm{ES}\Big(\widehat{P}^R_{x_i,\theta}, r_i\Big) = \sum_{j\in J_i} w^{(\theta)}_{ij}\|r_j - r_i\|_2 - \frac{1}{2}\sum_{j,k\in J_i} w^{(\theta)}_{ij} w^{(\theta)}_{ik}\|r_j - r_k\|_2,$$

which is exactly $\ell_i(\theta)$. □

Proof of Proposition 1. Conditional on X, $R \sim P^R_X$, while P^R_Z is fixed because $Z = h_\theta(X)$. Applying the energy-score identity conditionally on X gives

$$\mathbb{E}\big[\mathrm{ES}(P^R_Z, R) - \mathrm{ES}(P^R_X, R) \mid X\big] = \frac{1}{2}\mathcal{E}(P^R_X, P^R_Z).$$

Taking expectations over X proves the stated identity. Since energy distance is nonnegative and vanishes exactly when its arguments coincide, equality of the risks is equivalent to $P^R_X = P^R_Z$ almost surely, which is precisely uncertainty sufficiency. □

For the following result, let U denote the variable used for localization and use the kernel-weighted estimator in (3), with weights proportional to $\exp(-d_U(U_j, u)/\tau)$. Define the energy metric

$$d_{\mathrm{E}}(P, Q) := \left(\frac{\mathcal{E}(P, Q)}{2}\right)^{1/2}, \tag{8}$$

so that $d^2_{\mathrm{E}}(P, Q)$ equals the excess expected energy-score risk incurred by using Q when the true law is P.

Proposition 6 (Penalty from uncertainty-irrelevant dimensions)**.** *Let* $S = g(X) \in \mathbb{R}^d$ *be uncertainty-sufficient, and let* $U = (S, N) \in \mathbb{R}^p$*, where* $N = n(X) \in \mathbb{R}^{p-d}$*,* $p > d$*, and* $R \perp N \mid S$*. Fix a query with coordinates* $s_0 \in \mathrm{supp}(P_S)$ *and* $u_0 = (s_0, n_0) \in \mathrm{supp}(P_U)$*. Assume that, in neighborhoods of these points, the retrieval geometries are regular and equivalent to Euclidean geometry, and that* S *and* U *have continuous densities bounded above and below by positive constants. Assume also that* $P^R_s := \mathcal{L}(R \mid S = s)$ *is locally* β*-Hölder in energy metric: for some* $r > 0$*,* $d_{\mathrm{E}}(P^R_s, P^R_{s_0}) \le L\|s - s_0\|^\beta$*, for all* $s \in \mathrm{supp}(P_S) \cap B(s_0, r)$*, with* $0 < \beta \le 1$*, and that* $\mathcal{R}$ *is bounded.*

For the kernel estimator in (3)*, assume the isotropic kernel is bounded, nonnegative, nonzero near the origin, and has finite moments sufficient for the expressions below. Let* $\mathfrak{P}_{\beta,L}$ *denote the resulting class of distributions, and define* $\mathfrak{R}_d(\tau) := \sup_{P\in\mathfrak{P}_{\beta,L}} \mathbb{E}_P[\mathcal{E}(\widehat{P}^R_{S,\tau}, P^R_{s_0})]$ *and* $\mathfrak{R}_p(\tau) := \sup_{P\in\mathfrak{P}_{\beta,L}} \mathbb{E}_P[\mathcal{E}(\widehat{P}^R_{U,\tau}, P^R_{s_0})]$*. Then, as* $\tau \to 0$ *and* $m\tau^p \to \infty$*, where* $f \asymp g$ *means* $cg \le f \le Cg$ *for some constants* $0 < c \le C < \infty$*,*

$$\mathfrak{R}_d(\tau) \asymp \tau^{2\beta} + \frac{1}{m\tau^d}, \qquad \mathfrak{R}_p(\tau) \asymp \tau^{2\beta} + \frac{1}{m\tau^p}. \tag{9}$$

If $\tau^\star_d \in \arg\min_\tau \mathfrak{R}_d(\tau)$ *and* $\tau^\star_p \in \arg\min_\tau \mathfrak{R}_p(\tau)$*, then* $\tau^\star_d \asymp m^{-1/(2\beta+d)}$ *and* $\tau^\star_p \asymp m^{-1/(2\beta+p)}$*, and*

$$\mathfrak{R}_d(\tau^\star_d) \asymp m^{-2\beta/(2\beta+d)}, \qquad \mathfrak{R}_p(\tau^\star_p) \asymp m^{-2\beta/(2\beta+p)}. \tag{10}$$

Thus, although N *contains no additional information about the residual law, allowing it to vary under the retrieval metric slows the worst-case localization rate whenever* $p > d$*.*

Proof. We prove the result for a generic localization variable V of local dimension $s \in \{d, p\}$ and then substitute $V = S$ and $V = U$.

Let $\mu(s) := \mathbb{E}[\phi(R) \mid S = s]$ denote the conditional mean embedding associated with energy distance, so that $d_{\mathrm{E}}(P, Q) = \|\mu(P) - \mu(Q)\|_{\mathcal{H}}$ and $\mathcal{E}(P, Q) = 2\|\mu(P) - \mu(Q)\|_{\mathcal{H}}^2$. Writing the weighted empirical embedding as $\widehat{\mu}_{V,\tau} := \sum_i w_i^{V,\tau} \phi(R_i)$, we have

$$\widehat{\mu}_{V,\tau} - \mu(s_0) = \sum_i w_i^{V,\tau} \big(\phi(R_i) - \mu(S_i)\big) + \sum_i w_i^{V,\tau} \big(\mu(S_i) - \mu(s_0)\big).$$

Conditional on the reference locations, the first term has mean zero, so the cross term vanishes after expectation. Hence the expected squared embedding error separates into a finite-reference variance term and a squared localization-bias term.

For the bias, kernel localization at scale τ assigns asymptotically all effective weight to points whose retrieval distance from the query is $O(\tau)$. Local regularity of the retrieval geometry therefore implies $\|S_i - s_0\| = O(\tau)$ for effectively weighted samples. By the Hölder assumption, $\|\mu(S_i) - \mu(s_0)\|_{\mathcal{H}} = O(\tau^\beta)$, and thus the squared bias is $O(\tau^{2\beta})$.

For the variance, boundedness of $\mathcal{R}$ implies $\mathbb{E}[\|\phi(R) - \mu(S)\|_{\mathcal{H}}^2 \mid S] \le C$ for some finite C. Conditional independence therefore gives

$$\mathbb{E}\left[\left\|\sum_i w_i^{V,\tau} \big(\phi(R_i) - \mu(S_i)\big)\right\|_{\mathcal{H}}^2 \,\middle|\, V_1, \ldots, V_m\right] \le C \sum_i (w_i^{V,\tau})^2.$$

Because the local density of V is bounded above and below, standard kernel-sum concentration gives $\sum_i (w_i^{V,\tau})^2 = \Theta_p((m\tau^s)^{-1})$ and therefore an expected variance of order $(m\tau^s)^{-1}$. This proves the upper bound

$$\sup_{P \in \mathfrak{P}_{\beta,L}} \mathbb{E}_P[\mathcal{E}(\widehat{P}_{V,\tau}^R, P_{s_0}^R)] = O\left(\tau^{2\beta} + \frac{1}{m\tau^s}\right).$$

It remains to establish matching lower orders. For the variance term, restrict $\mathfrak{P}_{\beta,L}$ to a locally constant, nondegenerate two-point residual distribution. The localization bias then vanishes, while the conditional embedding variance is bounded below by a positive constant. The same kernel-weight calculation gives

$$\sup_{P \in \mathfrak{P}_{\beta,L}} \mathbb{E}_P[\mathcal{E}(\widehat{P}_{V,\tau}^R, P_{s_0}^R)] = \Omega\left(\frac{1}{m\tau^s}\right).$$

For the bias term, again restrict to residuals supported on two distinct points, but let the probability of one point vary locally as $\pi(s) = \pi_0 + c\|s - s_0\|^\beta$, with $c > 0$ chosen small enough that $\pi(s) \in (0, 1)$. This family lies in the assumed Hölder class. For two-point laws, energy distance is a positive constant times the squared difference in their probabilities. Local kernel averaging at scale τ changes $\pi(s_0)$ by order τ^β, because the local density is positive and the kernel assigns nonvanishing mass to points at distance proportional to τ. Hence

$$\sup_{P \in \mathfrak{P}_{\beta,L}} \mathbb{E}_P[\mathcal{E}(\widehat{P}_{V,\tau}^R, P_{s_0}^R)] = \Omega(\tau^{2\beta}).$$

Combining the two lower bounds and using $\max\{a, b\} \ge (a + b)/2$ yields

$$\sup_{P \in \mathfrak{P}_{\beta,L}} \mathbb{E}_P[\mathcal{E}(\widehat{P}_{V,\tau}^R, P_{s_0}^R)] \asymp \tau^{2\beta} + \frac{1}{m\tau^s}.$$

Taking $s = d$ and $s = p$ proves (9). Finally, minimizing $\tau^{2\beta} + (m\tau^s)^{-1}$ gives $\tau_s^\star \asymp m^{-1/(2\beta+s)}$ and optimized risk $m^{-2\beta/(2\beta+s)}$, yielding (10). □

Interpretation. The p-dimensional penalty applies only when the additional coordinates introduce genuine local variation under the retrieval metric. If a p-dimensional representation is merely a redundant embedding of a d-dimensional manifold, its local probability mass still scales according to dimension d, and the p-dimensional penalty does not apply. Thus, the relevant quantity is the intrinsic localization dimension rather than the number of stored coordinates.

To see the statistical mechanism, define the effective local support

$$N_{\text{eff}}(u,\tau) := \frac{1}{\sum_j w_j(u)^2}. \tag{11}$$

At a fixed localization scale, uncertainty-irrelevant directions can separate reference samples under the retrieval metric without improving approximation of the conditional residual law, thereby reducing N_{eff}. Maintaining the same effective support then requires broader localization, which increases bias. An ideal uncertainty-relevant representation collapses such nuisance directions, whereas representation-learning error may leave variation along them and increase the empirical localization dimension.

This benefit must be balanced against uncertainty sufficiency. Excessive compression can discard information that genuinely changes P_x^R, incurring the representation error characterized in Prop. 1. The desired representation therefore preserves uncertainty-relevant variation while suppressing nuisance variation that makes residual-law retrieval statistically harder.

Theorem 7 (Finite-sample marginal validity)**.** *Let $\mathcal{D}_0$ denote all data and randomness used to define $\widehat{S}$. Assume that, conditional on $\mathcal{D}_0$, $(X_1,Y_1),\dots,(X_n,Y_n),(X_{\text{new}},Y_{\text{new}})$ are exchangeable. Then*

$$\mathbb{P}\Big(Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}}) \mid \mathcal{D}_0\Big) \geq 1-\alpha, \tag{12}$$

$$\mathbb{P}\Big(Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}})\Big) \geq 1-\alpha. \tag{13}$$

Proof. Conditional on $\mathcal{D}_0$, the score function $\widehat{S} : \mathcal{X}\times\mathcal{Y} \to [0,\infty]$ is fixed. Since $(X_1,Y_1),\dots,(X_n,Y_n),(X_{\text{new}},Y_{\text{new}})$ are exchangeable conditional on $\mathcal{D}_0$, applying the same fixed function $\widehat{S}$ to each observation implies that the scores $S_1,\dots,S_n,S_{\text{new}}$, where $S_{\text{new}} = \widehat{S}(X_{\text{new}},Y_{\text{new}})$, are also exchangeable conditional on $\mathcal{D}_0$.

Let $k = \lceil (n+1)(1-\alpha)\rceil$. By exchangeability, the rank of S_{new} among the $n+1$ scores is uniform up to ties. Hence the split-conformal order statistic satisfies

$$\mathbb{P}\big(S_{\text{new}} \leq S_{(k)} \mid \mathcal{D}_0\big) \geq \frac{k}{n+1} \geq 1-\alpha. \tag{14}$$

Here $S_{(k)}$ denotes the kth order statistic of the calibration scores, with the convention $S_{(n+1)} = \infty$.

By definition of the conformal prediction set, $\widehat{C}_\alpha(x) = \{y \in \mathcal{Y} : \widehat{S}(x,y) \leq S_{(k)}\}$, so the event $S_{\text{new}} \leq S_{(k)}$ is exactly $Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}})$. Equivalently, since $\widehat{S}(x,y) \leq a$ if and only if $T(x,y) \in \widehat{G}_a(x)$, this event is the same as $T(X_{\text{new}},Y_{\text{new}}) \in \widehat{G}_{S_{(k)}}(X_{\text{new}})$. Therefore,

$$\mathbb{P}\Big(Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}}) \mid \mathcal{D}_0\Big) \geq 1-\alpha, \tag{15}$$

which proves (12).

Finally, by the law of total probability, $\mathbb{P}(A) = \mathbb{E}[\mathbb{P}(A \mid \mathcal{D}_0)]$. Applying this to the event $A = \{Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}})\}$ and using the conditional bound above gives

$$\mathbb{P}\Big(Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}})\Big) = \mathbb{E}\Big[\mathbb{P}\Big(Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}}) \mid \mathcal{D}_0\Big)\Big] \geq 1-\alpha, \tag{16}$$

proving (13). □

Assumption 8 (Oracle common-coordinate representation)**.** *There exists a probability law Q_0 on a canonical residual-coordinate space $\mathcal{C}$ and measurable maps $\Phi_x : \mathcal{R} \to \mathcal{C}$, indexed by $x \in \mathcal{X}_0$, such that $(\Phi_x)_\# P_x^R = Q_0$ for all $x \in \mathcal{X}_0$ ($\#$ denotes pushforward).*

Fix a nested family $\{K_a\}_{a\geq 0}$ in $\mathcal{C}$ and define $\rho(u) := \inf\{a \geq 0 : u \in K_a\}$. The induced oracle geometry and score are

$$G_a^\star(x) := \{r \in \mathcal{R} : \Phi_x(r) \in K_a\}, \qquad S^\star(x,y) := \rho(\Phi_x(T(x,y))),$$

and let $F_0(a) := Q_0(K_a)$. This construction is used only for analysis; FacCP does not estimate Φ_x explicitly.

Theorem 9 (Oracle pivotality)**.** *Under Assumption 8, if $U \sim Q_0$, then $S^\star(X,Y) \mid X = x \overset{d}{=} \rho(U)$ for all $x \in \mathcal{X}_0$. Hence $S^\star$ is pivotal with respect to X.*

Proof. Conditional on $X = x$, Assumption 8 gives $\Phi_x(R) \sim Q_0$. Applying the same measurable functional ρ at every query gives $\rho(\Phi_x(R)) \overset{d}{=} \rho(U)$, whose distribution is independent of x. □

Proof of Theorem 4. By the score-set equivalence (i.e., $\widehat{S}(x,y) \le a$ is equivalent to $T(x,y) \in \widehat{G}_a(x)$), $F_x(a) = P_x^R(\widehat{G}_a(x))$, while we also have $F_0(a) = P_x^R(G_a^\star(x))$ for all $x \in \mathcal{X}_0$ and for all $a \ge 0$. For any probability measure P and measurable sets A, B, $|P(A) - P(B)| \le P(A \triangle B)$. Hence Assumption 3 gives $|F_x(a) - F_0(a)| \le \varepsilon_{\mathrm{geom}}$ uniformly in $x \in \mathcal{X}_0$ and $a \ge 0$.

Averaging $F_x(a)$ over X gives $F(a)$, so $|F(a) - F_0(a)| \le \varepsilon_{\mathrm{geom}}$. Therefore,

$$|F_x(a) - F(a)| \le |F_x(a) - F_0(a)| + |F(a) - F_0(a)| \le 2\varepsilon_{\mathrm{geom}}, \tag{17}$$

and taking the supremum proves the result. □

Corollary 10 (Pivotality controls population local coverage)**.** *Suppose the score $\widehat{S}$ has pivotality defect $\Delta_{\mathrm{piv}} = \sup_{x \in \mathcal{X}_0,\, a \ge 0} |F_x(a) - F(a)|$. Then, for any threshold a,*

$$\sup_{x \in \mathcal{X}_0} |F_x(a) - (1-\alpha)| \le \Delta_{\mathrm{piv}} + |F(a) - (1-\alpha)|\,. \tag{18}$$

In particular, let $q_\alpha := F^{-1}(1-\alpha) := \inf\{a : F(a) \ge 1-\alpha\}$, where F^{-1} denotes the generalized inverse, and suppose $F(q_\alpha) = 1-\alpha$. Then

$$\sup_{x \in \mathcal{X}_0} \left| \mathbb{P}\Big(\widehat{S}(X,Y) \le q_\alpha \mid X = x\Big) - (1-\alpha) \right| \le \Delta_{\mathrm{piv}}. \tag{19}$$

Thus, if the exact marginal population quantile were known, the pivotality defect alone bounds the worst-case local coverage error.

Proof. For any x and threshold a, the triangle inequality gives

$$|F_x(a) - (1-\alpha)| \le |F_x(a) - F(a)| + |F(a) - (1-\alpha)|\,.$$

By the definition of Δ_{piv}, $|F_x(a) - F(a)| \le \Delta_{\mathrm{piv}}$ uniformly over $x \in \mathcal{X}_0$ and $a \ge 0$. Taking the supremum over $x \in \mathcal{X}_0$ therefore yields

$$\sup_{x \in \mathcal{X}_0} |F_x(a) - (1-\alpha)| \le \Delta_{\mathrm{piv}} + |F(a) - (1-\alpha)|\,,$$

which proves (18).

Now set $a = q_\alpha$. Since $F(q_\alpha) = 1-\alpha$, the second term vanishes, giving

$$\sup_{x \in \mathcal{X}_0} |F_x(q_\alpha) - (1-\alpha)| \le \Delta_{\mathrm{piv}}.$$

Finally, by definition, $F_x(q_\alpha) = \mathbb{P}(\widehat{S}(X,Y) \le q_\alpha \mid X = x)$, which gives (7). □

α-local pivotality. For $\eta \ge 0$, define the near-quantile region and its pivotality defect by

$$\mathcal{T}_{\alpha,\eta} := \{a :\ |F(a) - (1-\alpha)| \le \eta\}, \qquad \Delta_{\alpha,\eta} := \sup_{x \in \mathcal{X}_0} \sup_{a \in \mathcal{T}_{\alpha,\eta}} |F_x(a) - F(a)|.$$

Thus, for any $a \in \mathcal{T}_{\alpha,\eta}$, Cor. 10 gives

$$\sup_{x \in \mathcal{X}_0} |F_x(a) - (1-\alpha)| \le \Delta_{\alpha,\eta} + \eta.$$

Empirically, we use

$$\widehat{\Delta}_{\alpha,\eta_n} := \max_g \max_{a:\, |\widehat{F}_{\mathrm{cal}}(a) - (1-\alpha)| \le \eta_n} |\widehat{F}_g(a) - \widehat{F}(a)|, \qquad \eta_n := \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2}{n},$$

where g ranges over reference-fit query-difficulty bins. This is a held-out finite-bin diagnostic, not a uniform conditional-coverage certificate.

Theorem 11 (Local coverage under global calibration)**.** *Assume that, conditional on $\mathcal{D}_0$, the calibration observations are i.i.d. from the target distribution and that the marginal score distribution F is continuous. Let $k = \lceil (n+1)(1-\alpha) \rceil \le n$ and $\widehat{q}_\alpha = S_{(k)}$. Then, with probability at least $1-\delta$ over $\mathcal{D}_{\mathrm{cal}}$,*

$$\sup_{x \in \mathcal{X}_0} \left| \mathbb{P}\Big(Y_{\mathrm{new}} \in \widehat{C}_\alpha(x) \mid X_{\mathrm{new}} = x,\, \mathcal{D}_0,\, \mathcal{D}_{\mathrm{cal}} \Big) - (1-\alpha) \right| \le \Delta_{\mathrm{piv}} + \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2}{n}. \quad (20)$$

Proof. Let $\widehat{F}_n(a) := n^{-1} \sum_{i=1}^n \mathbf{1}\{S_i \le a\}$. By the Dvoretzky–Kiefer–Wolfowitz inequality, with probability at least $1-\delta$,

$$\sup_a \left| \widehat{F}_n(a) - F(a) \right| \le \varepsilon_n(\delta), \qquad \varepsilon_n(\delta) := \sqrt{\frac{\log(2/\delta)}{2n}}. \quad (21)$$

Continuity of F implies that the calibration scores are almost surely distinct, so $\widehat{F}_n(\widehat{q}_\alpha) = k/n$. By the definition of k, $|k/n - (1-\alpha)| \le 2/n$. Hence, on the DKW event,

$$|F(\widehat{q}_\alpha) - (1-\alpha)| \le \varepsilon_n(\delta) + \frac{2}{n}. \quad (22)$$

For every $x \in \mathcal{X}_0$, the definition of Δ_{piv} gives

$$|F_x(\widehat{q}_\alpha) - (1-\alpha)| \le \Delta_{\mathrm{piv}} + \varepsilon_n(\delta) + \frac{2}{n}. \quad (23)$$

Finally, by score-set equivalence (i.e., $\widehat{S}(x,y) \le a$ is equivalent to $T(x,y) \in \widehat{G}_a(x)$), $F_x(\widehat{q}_\alpha)$ is exactly the conditional coverage probability of $\widehat{C}_\alpha(x)$ given $\mathcal{D}_0$ and $\mathcal{D}_{\mathrm{cal}}$. □

Corollary 12 (Covariate-shift transfer)**.** *Let the calibration sample be i.i.d. from P^{s}, let $\widehat{q}_\alpha$ be the resulting unweighted split-conformal threshold, and assume the source marginal score distribution F is continuous. Then, with probability at least $1-\delta$ over the source calibration sample,*

$$\sup_{x \in \mathcal{X}_0} \left| P^{\mathrm{t}} \left(Y_{\mathrm{new}} \in \widehat{C}_\alpha(x) \mid X_{\mathrm{new}} = x,\, \mathcal{D}_0, \mathcal{D}_{\mathrm{cal}} \right) - (1-\alpha) \right| \le \Delta_{\mathrm{piv}} + \sqrt{\tfrac{\log(2/\delta)}{2n}} + \tfrac{2}{n}, \quad (24)$$

where the supremum is over $x \in \mathcal{X}_0$. Consequently, the same bound holds for target marginal coverage. If Assumption 3 holds, then $\Delta_{\mathrm{piv}} \le 2\varepsilon_{\mathrm{geom}}$, so target-domain coverage is controlled by the same geometry and global-calibration terms as in the unshifted setting.

Proof. For the proof, write

$$F_x^{\mathrm{s}}(a) = P^{\mathrm{s}} \left(\widehat{S}(x,Y) \le a \mid X = x \right), \qquad F_x^{\mathrm{t}}(a) = P^{\mathrm{t}} \left(\widehat{S}(x,Y) \le a \mid X = x \right).$$

Conditional on $\mathcal{D}_0$, the learned score $\widehat{S}$ is fixed. Since covariate shift preserves the conditional distribution of Y given $X = x$,

$$F_x^{\mathrm{t}}(a) = F_x^{\mathrm{s}}(a)$$

for every $x \in \mathcal{X}_0$ and a.

Let F^{s} denote the source marginal score distribution. By the same finite-calibration argument as in Theorem 11, with probability at least $1-\delta$,

$$|F^{\mathrm{s}}(\widehat{q}_\alpha) - (1-\alpha)| \le \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2}{n}.$$

Therefore, for every $x \in \mathcal{X}_0$,

$$\begin{aligned} \left| F_x^{\mathrm{t}}(\widehat{q}_\alpha) - (1-\alpha) \right| &= |F_x^{\mathrm{s}}(\widehat{q}_\alpha) - (1-\alpha)| \\ &\le |F_x^{\mathrm{s}}(\widehat{q}_\alpha) - F^{\mathrm{s}}(\widehat{q}_\alpha)| + |F^{\mathrm{s}}(\widehat{q}_\alpha) - (1-\alpha)| \\ &\le \Delta_{\mathrm{piv}} + \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2}{n}. \end{aligned}$$

Taking the supremum over $x \in \mathcal{X}_0$ gives the stated conditional bound. Averaging over P_X^{t} gives the target marginal bound. □

This perspective is related to Gibbs et al. (2025), who formulate conditional validity as coverage over classes of covariate shifts and calibrate against specified classes of reweightings. FacCP takes a complementary route: rather than adapting calibration to a target shift, it seeks an approximately pivotal score whose conditional law is invariant in x. When this structural property holds over the target-relevant region, the same unweighted source threshold transfers without target-specific density-ratio estimation; when it fails, shift-aware calibration remains necessary (Tibshirani et al., 2019; Barber et al., 2023; Joshi et al., 2026).

B Additional Related Work

Interpreting Table 2. Table 2 compares methods by where adaptivity enters and how uncertainty is represented and calibrated. "Query adaptive" indicates that the resulting score or prediction set varies with the query, while "Multivariate" indicates support for genuinely multivariate outputs. "Full conditional law" means that the method models, embeds, or retrieves an entire conditional output/residual distribution rather than only a scale, quantile, moment, or scalar score distribution. "Uncertainty metric" indicates that a covariate representation is learned specifically so that proximity reflects similarity in uncertainty. "Adaptive geometry" means that the multivariate set shape itself can vary across queries rather than merely rescaling or re-indexing a fixed geometry. "Global calibration" indicates that a single scalar quantile is estimated from the full calibration sample. "Reusable geometry" means that the learned uncertainty representation can be paired with different downstream set constructors without retraining. "Local coverage theory" denotes an explicit local or conditional coverage result beyond ordinary marginal validity. Checkmarks denote properties that are explicit components of the primary method analyzed; crosses indicate only that the property is not part of that primary construction.

The rows are grouped by the main source of adaptivity: scalar-score or predictor adaptation, localized or group-conditional calibration, conditional-law modeling, and multivariate geometry or transformations. These groups highlight the central tradeoff: scalar and localized methods operate on lower-dimensional uncertainty summaries or calibration rules, whereas distributional and multivariate methods retain richer uncertainty structure. FacCP instead learns a query-space metric for retrieving the residual law, uses that law to construct query-dependent geometry before scalarization, and retains a single globally pooled conformal calibration step.

Table 2: Comparison of representative adaptive conformal prediction methods. A checkmark indicates that a property is part of the method's primary construction; × indicates otherwise.

Method	Query adaptive	Multi-variate	Full conditional law	Uncertainty metric	Adaptive geometry	Global calibration	Reusable geometry	Local coverage theory
Scalar-score / predictor adaptation								
Normalized nonconformity (Papadopoulos et al., 2008) (Papadopoulos & Haralambous, 2011)	✓	×	×	×	×	✓	×	×
CQR (Romano et al., 2019)	✓	×	×	×	×	✓	×	×
Self-supervised adaptive CP (Seedat et al., 2023)	✓	×	×	×	×	✓	×	×
Regression adjustment (Gao et al., 2024)	✓	×	×	×	×	✓	×	✓
Split-Jackknife+ adaptive scores (Deutschmann et al., 2024)	✓	×	×	×	×	×	×	✓
Rectified CP (Plassier et al., 2025a)	✓	✓	×	×	×	✓	×	✓
PIT-CP (Laplante, 2026)	✓	✓	×	×	×	✓	×	✓
Localized / group-conditional calibration or scores								
Localized CP / RLCP (Guan, 2023; Hore & Barber, 2025)	✓	×	×	×	×	×	×	✓
Split Localized CP (Han et al., 2022)	✓	×	×	×	×	✓	×	✓
PLCP (Kiyani et al., 2024)	✓	×	×	✓	×	×	×	✓
Clustered class-conditional CP (Ding et al., 2023)	×	×	×	×	×	×	×	✓
Conditional-law / density / depth models								
Distributional CP (Chernozhukov et al., 2021)	✓	×	✓	×	×	✓	×	✓
SPICE (Diamant et al., 2024)	✓	×	✓	×	×	✓	×	✓
PCP (Wang et al., 2023)	✓	✓	✓	×	✓	✓	×	×
CP^2 (Plassier et al., 2025b)	✓	✓	✓	×	✓	✓	✓	✓
Kernel-depth CP (Matabuena et al., 2024)	✓	✓	✓	×	✓	✓	×	✓
Joint-density TPP regions (Dheur et al., 2024)	✓	✓	✓	×	✓	✓	×	×
Multivariate geometry / transformations								
Adaptive ellipsoids (Messoudi et al., 2022)	✓	✓	×	×	✓	✓	×	×
Multi-output representation QR (Feldman et al., 2023)	✓	✓	×	×	✓	✓	×	×
CONTRA / JAPAN (Fang et al., 2025; English & Lippert, 2026)	✓	✓	✓	×	✓	✓	×	×
OT-based CP (Thurin et al., 2025)	✓	✓	×	×	✓	✓	×	✓
FacCP (ours)	✓	✓	✓	✓	✓	✓	✓	✓

This appendix expands Table 2. The table is intended to compare *where adaptivity is learned and where calibration occurs*, rather than to rank methods by generality. A × means only that a property is not part of the primary construction analyzed in the cited work. Several references below are not included as table rows because they primarily establish conceptual foundations, diagnostics, or

task-specific conditional guarantees rather than an alternative architecture for adaptive multivariate regression.

Conditional validity as a distinct target. The distinction between marginal and conditional validity predates recent adaptive conformal methods. Vovk (2012) systematically studies several notions of conditional validity for inductive conformal predictors and modifications designed to obtain them. This work is useful for situating our objective: FacCP does not claim distribution-free exact pointwise conditional validity, which is impossible in general without additional structure, but seeks a score whose conditional law varies little with the query so that global calibration remains locally meaningful. Cauchois et al. (2021) make a related point in multiclass and multilabel prediction, where ordinary marginally valid conformal sets can allocate errors unevenly between easy and difficult examples. They use quantile-regression ideas to obtain asymptotically consistent conditional behavior while preserving marginal validity. These works motivate conditional adaptivity, but do not address the problem central to FacCP of learning a reusable multivariate residual-law representation whose geometry is constructed before scalar conformal calibration.

Early normalized nonconformity scores. Adaptive nonconformity scores have a long history in conformal regression. Papadopoulos et al. (2008) observe that an unnormalized regression residual tends to produce prediction intervals of similar width across examples and introduce normalized nonconformity measures whose widths depend on an estimate of local prediction difficulty. Papadopoulos & Haralambous (2011) extend this principle to neural-network regression and produce intervals whose widths vary with the expected difficulty of each test example. These methods are important predecessors of modern learned-score approaches: they absorb heteroscedasticity into the scalar score and retain the usual conformal calibration mechanism. Their limitation relative to FacCP is also representative of scalar normalization more broadly. A single local scale can adapt interval width when residual laws are approximately related by rescaling, but it does not model changes in multivariate orientation, dependence, skewness, or modality. FacCP replaces the scalar difficulty model with an estimate of the full conditional residual law and therefore allows the geometry itself to change with the query.

Training the predictor for conditionally homogeneous scores. Gao et al. (2024) take a particularly relevant approach: rather than modifying only the post-hoc calibration rule, they train the regression function to improve conditional coverage after split conformal calibration. Their analysis observes that conditional miscoverage is driven by disagreement between the marginal and conditional distributions of the nonconformity score, and uses the Kolmogorov–Smirnov distance between these score laws as a regularization target. This is closely aligned with the pivotality principle used in our analysis.

The location of the adaptation, however, is different. Gao et al. modify the *point predictor* so that a chosen scalar score becomes more homogeneous across x. This can be attractive when the predictor is still trainable and the desired set family is adequately represented by that score, but it couples predictive accuracy and uncertainty calibration in a single training objective and does not by itself recover missing multivariate residual geometry. FacCP treats $\widehat{f}$ as fixed and learns a separate uncertainty representation targeted at P_x^R. This preserves the predictor, permits uncertainty factors to differ from prediction-relevant factors, and allows the resulting residual law to drive multiple downstream geometries. Thus both methods seek score-distribution homogeneity, but Gao et al. act through the predictive model whereas FacCP acts through query-dependent residual geometry.

Pivotal-score post-processing: PIT-CP. PIT-CP is the closest prior work to the pivotality viewpoint in Sec. 4.3. It constructs a pivotalized score by applying an estimate of the conditional CDF of an arbitrary base nonconformity score (Laplante, 2026). This reduces the learned conditional object to one dimension and preserves the geometry of the original score family. That is a statistical advantage when the base score already captures the relevant uncertainty geometry.

The same preservation creates the principal distinction from FacCP. PIT-CP can correct how a fixed nested geometry is indexed across queries, but cannot introduce residual structure that the base score discards. For example, transforming the conditional CDF of a spherical residual norm does not by itself recover a query-dependent covariance orientation or multimodal residual shape. FacCP instead learns $\widehat{P}_x^R$ before scalarization and constructs $\widehat{G}_t(x)$ from that law. Its geometry-to-pivotality

analysis therefore addresses the upstream question of how an adaptive multivariate score becomes approximately pivotal.

Rectifying conformity scores. Rectified CP learns conditional score quantiles and transforms a base conformity score so that its relevant conditional quantile is approximately invariant across covariates (Plassier et al., 2025a). Like PIT-CP, it offers an efficient scalar route to improved conditional coverage and retains standard pooled split-conformal calibration. Its conditional model need only describe the chosen score rather than the full response or residual distribution. Relative to FacCP, this reduction is both its strength and its restriction: score rectification inherits the representational choices of the base score, whereas FacCP learns the uncertainty object from which different score geometries can subsequently be constructed.

Adaptive Split-Jackknife+ scores. Deutschmann et al. (2024) adapt nonconformity scores using held-out data and, in their principal regression construction, normalize residuals using an estimate of local prediction error. Such scalar correction can be highly effective when heterogeneity is well described by local scale. FacCP is aimed at the more general setting where uncertainty can vary in shape and dependence as well as scale. It also isolates adaptation to an auxiliary residual-law sample, keeping the final calibration sample dedicated to one pooled scalar quantile.

Split Localized Conformal Prediction. Split Localized CP uses kernel estimates of local conditional information to modify the nonconformity score before split-conformal calibration (Han et al., 2022). This is closer to FacCP than methods that localize the final calibration quantile, because localization is partly absorbed into the score itself. However, its effectiveness depends on the chosen feature space and kernel locality. FacCP makes that representation problem an explicit learning objective: proximity under h_θ is trained according to the residual distribution retrieved by the resulting neighbors. Its uncertainty-sufficiency and intrinsic-dimension analyses further formalize why suppressing uncertainty-irrelevant directions can improve local sample support.

Conditional-law modeling and CP^2. CP^2 combines conformal calibration with an estimate of the full conditional response law $P_{Y|X}$ and derives approximate conditional-coverage bounds in terms of error in that conditional model (Plassier et al., 2025b). This makes it one of the closest distributional alternatives to FacCP. The principal distinction is the modeled object: CP^2 asks a conditional model to represent the response distribution, whereas FacCP assumes a fixed predictor and estimates only its residual law. When the point predictor already explains substantial response variation, P_x^R can be statistically simpler than $P_{Y|X=x}$. FacCP additionally learns the coordinates in which residual-law localization should occur and separates representation, localization, finite-reference, geometry, pivotality, and final calibration errors.

Kernel-depth prediction regions for structured responses. Matabuena et al. (2024) provide an especially relevant comparison for complex response spaces. They use conditional depth measures and conditional kernel mean embeddings to define prediction and tolerance regions for responses in separable Hilbert spaces, focusing empirically on functional data. With characteristic kernels, a kernel mean embedding can uniquely represent a probability law, so this framework can retain rich distributional information rather than only low-order moments. They also develop a conformal variant with finite-sample marginal validity and study conditional and unconditional consistency.

This work shares FacCP's motivation of avoiding overly restrictive scalar uncertainty summaries, but the representation and modularity are different. Kernel-depth CP commits to a depth construction in a chosen response-space RKHS; the conditional embedding is estimated for the purpose of that depth-induced region. FacCP instead learns a *query-space* representation whose only target is residual-law retrieval, leaves the residuals in their original space, and separates retrieval from the downstream geometry constructor. Consequently, FacCP can reuse the same learned h_θ with ellipsoids, hyperboxes, zonotopes, or other set families, while kernel-depth CP provides a particularly powerful but more specific geometry for structured responses.

Joint density regions for structured multivariate outputs. Dheur et al. (2024) construct conformal joint prediction regions for neural marked temporal point processes, where the output combines a continuous event time with a categorical mark. Their sharper construction uses bivariate highest-density regions derived from the model's joint predictive density, thereby exploiting dependence

between output components rather than conformalizing each component independently. This is an important example of the efficiency gained by using the *joint* predictive law.

The method is nevertheless tied to a task-specific probabilistic model and to highest-density geometry induced by its joint density. FacCP similarly seeks to exploit joint residual structure, but does not require the base predictor to output a normalized predictive density. Instead it constructs an empirical residual law around the query and allows the downstream set representation to be chosen independently of the predictive model.

Localized and group-conditional calibration. Localized conformal prediction and randomized local weighting use query-dependent calibration weights (Guan, 2023; Hore & Barber, 2025); PLCP learns features that support localization or partitioning (Kiyani et al., 2024); and broader conditional-guarantee frameworks target structured classes of conditional variation (Gibbs et al., 2025). These methods can directly enforce behavior on the target groups or neighborhoods, but query-dependent quantiles may have substantially less effective support than the entire calibration sample.

The classification literature illustrates the same support tradeoff particularly clearly. Ding et al. (2023) study class-conditional coverage when there are many classes and few labeled calibration examples per class. Their clustered conformal method pools classes with similar conformal-score distributions before calibrating at the cluster level, improving the data available for class-conditional calibration. The underlying idea is closely related to FacCP's motivation—pool observations whose uncertainty behavior is similar—but the implementation occurs at the level of discrete output classes and cluster-specific calibration thresholds. FacCP instead learns similarity between *queries* according to the complete residual law, uses the pooling only for auxiliary law estimation, and retains one global conformal threshold.

Conditional-coverage diagnostics. Conditional coverage also raises an evaluation problem: empirical marginal coverage can hide substantial local miscalibration, while naive covariate binning can be statistically weak in moderate or high dimensions. Cauchois et al. (2021) emphasize this issue in multiclass and multilabel prediction and introduce empirical validation procedures for more robust coverage assessment. More recently, Braun et al. (2025) formulate conditional-coverage diagnostics as a classification problem and propose excess-risk metrics designed to detect deviations from the target coverage with greater statistical power than simple binning-based measures.

These papers are diagnostic rather than competing set constructors, so we do not include them as separate rows in Table 2. They are directly relevant to our experimental evaluation, however: because FacCP's theory predicts that improved geometry reduces pivotality defect and hence local coverage error, strong conditional-coverage diagnostics provide a complementary empirical test of whether this mechanism actually holds beyond coarse covariate partitions.

Score values versus score ordering. Hanselle et al. (2025) emphasize that conformal prediction is invariant to strictly monotone transformations of the nonconformity score: for ordinary conformal validity, only the induced ordering matters. They use this observation to formulate conformal prediction through pairwise preferences rather than explicit real-valued nonconformity scores. This perspective clarifies the role of our scalarization. The numerical scale of $\widehat{S}(x, y)$ is not intrinsically meaningful for marginal conformal validity; what matters for FacCP is the stronger objective that its conditional distribution be comparable across queries. Approximate pivotality therefore imposes structure beyond the score ordering needed for marginal validity, because the same globally calibrated score level is intended to have approximately common conditional probability content across x.

Conditional-density and generative approaches. COPS, distributional conformal prediction, Dist-/CD-split, HPD-split, and SPICE construct adaptive sets from conditional density or distribution estimates (Lei & Wasserman, 2014; Chernozhukov et al., 2021; Izbicki et al., 2020; 2022; Diamant et al., 2024). PCP instead forms potentially disconnected regions from conditional generative samples (Wang et al., 2023). These methods can capture richer uncertainty than scalar scale models, but require a conditional density, CDF, or generative sampler. FacCP represents $\widehat{P}_x^R$ directly as a weighted empirical law, avoiding the need to learn a normalized residual density and allowing multiple downstream summaries of the same retrieved residual distribution.

Multivariate geometry and learned transformations. Multivariate conformal methods construct adaptive ellipsoids (Messoudi et al., 2022), learn response representations (Feldman et al., 2023), optimize multimodal shape templates (Tumu et al., 2024), construct time-series ellipsoids (Xu et al., 2024), or replace scalar ranks with vector optimal-transport ranks (Thurin et al., 2025). CONTRA and JAPAN learn conditional normalizing-flow transformations or densities whose level sets yield flexible multivariate regions (Fang et al., 2025; English & Lippert, 2026). These methods demonstrate that expressive geometry can materially improve multivariate efficiency, but typically tie the learned uncertainty model to a particular output representation. FacCP makes the retrieved residual law the intermediate object so that the geometry may be changed independently of representation learning.

Covariate shift and nonexchangeability. Under covariate shift, weighted conformal prediction changes calibration weights according to source-to-target likelihood ratios (Tibshirani et al., 2019), while broader nonexchangeable methods and high-dimensional shift corrections alter the final conformal calibration rule (Barber et al., 2023; Joshi et al., 2026). FacCP takes a different route: if residual-law adaptation makes the score approximately pivotal over source- and target-relevant covariates, then covariate shift changes the frequency of queries but not their conditional score law, permitting reuse of the same unweighted global threshold. This structural advantage does not replace shift-aware calibration when pivotality fails.

Conformal geometry in control and reachability. Adaptive multivariate regions are especially relevant in robotics and control because their geometry directly affects downstream robust optimization. Conformalized system-level synthesis uses learned state–control-dependent covariance models and conformal calibration to construct robust reachable tubes (Srinivasan et al., 2026), while shape-template methods emphasize sets consumable by planning algorithms (Tumu et al., 2024). FacCP moves the learning target upstream: instead of committing during training to a particular covariance or reachable-set parameterization, it learns the conditional residual law from which multiple computationally convenient geometries can subsequently be constructed.

C Additional Experimental Details

C.1 Controlled Residual-Geometry Experiment

Data-generating process. Let $X = (X_1, X_2, X_3, N)$, where $X_1, X_2, X_3 \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[-1, 1]$ and $N \sim \mathcal{N}(0, I_{d_{\mathrm{nuis}}})$ is independent. Residuals follow

$$R \mid X = x \sim \mathcal{N}(0, \Sigma(x)), \qquad x = (x_1, x_2, x_3, n),$$

with

$$s(x) = \exp(0.65 x_1), \qquad a(x) = \exp\left\{ \frac{\log 8}{2}(x_2 + 1) \right\}, \qquad \theta(x) = \frac{\pi}{2} x_3,$$

and

$$\Sigma(x) = Q_{\theta(x)} \begin{bmatrix} s(x)^2 a(x) & 0 \\ 0 & s(x)^2 / a(x) \end{bmatrix} Q_{\theta(x)}^\top.$$

Thus the first three coordinates determine the scale, aspect ratio, and orientation of the conditional residual distribution, while the nuisance coordinates change X without changing P_x.

We use independent samples of size $50{,}000$ for representation training, $30{,}000$ for residual retrieval, $8{,}000$ for hyperparameter tuning, $10{,}000$ for conformal calibration, and $10{,}000$ for evaluation. The target miscoverage is 5%. All learned components and retrieval hyperparameters are fixed before conformal calibration.

Geometry constructions. Oracle geometry uses the exact conditional covariance $\Sigma(x)$ in the Mahalanobis score. Its population score distribution is invariant in x. Small nonzero empirical pivotality and coverage errors therefore arise from finite calibration and Monte Carlo evaluation.

Global geometry estimates one residual covariance from the full reference sample and uses it at every query. Only its final conformal radius changes, so it cannot represent query-dependent aspect ratio or orientation.

Privileged retrieval uses

$$\psi(x) = (\log s(x), \log a(x) \cos 2\theta(x), \log a(x) \sin 2\theta(x)) .$$

The doubled angle encodes the π-periodicity of ellipse orientation. These coordinates are computed from the known data-generating factors, but they do not provide $\Sigma(x)$. The method uses the same finite reference sample, kernel retrieval, and weighted covariance construction as the other retrieval methods. It therefore isolates the effect of localizing in coordinates that contain the relevant uncertainty factors without nuisance variation.

Raw-X retrieval uses the same kernel retrieval and covariance construction directly in the observed input space. When $d_{\text{nuis}} = 0$, X is low-dimensional and its three coordinates determine P_x. Increasing d_{nuis} adds coordinates that affect retrieval distances without changing the conditional residual distribution.

FacCP replaces the prescribed localization coordinates with a learned energy-score representation. The representation has dimension 12 and hidden width 64 and is trained for 150 epochs. The retrieved residuals define a weighted second moment, which is used as the local ellipsoidal geometry.

Covariance prediction uses a width-64 neural network to predict a positive-definite 2×2 covariance directly from X and is trained using Gaussian negative log likelihood. Since the conditional residual distribution in this experiment is Gaussian and the target sets are ellipsoidal, this baseline is well specified.

Retrieval and calibration. For each retrieval representation $z(x)$, reference coordinates are standardized using the reference sample, except for the learned representation whose metric is fixed by training. Retrieval weights are proportional to

$$w_j(x) \propto \exp\left(-\frac{\|\widetilde{z}(x) - \widetilde{z}(X_j)\|_2^2}{\tau}\right).$$

The resulting covariance is the weighted residual second moment with relative ridge 10^{-4}. The temperature τ is selected on the independent tuning split by Gaussian negative log likelihood among candidates with mean effective sample size at least 100. A single split-conformal radius is then estimated using all $10{,}000$ calibration observations.

α-local pivotality diagnostic. We use an empirical diagnostic from Cor. 10 rather than attempting to estimate the theoretical supremum over every x. For each of scale, aspect ratio, and orientation, we form 10 equal-frequency bins using boundaries estimated from the reference sample. Within each held-out evaluation bin, we compare the conditional empirical score CDF with the pooled evaluation CDF over calibration thresholds satisfying

$$\left|\widehat{F}_{\text{cal}}(t) - 0.95\right| \le \eta_n, \qquad \eta_n = \sqrt{\frac{\log(2/0.05)}{2n_{\text{cal}}}} + \frac{2}{n_{\text{cal}}} = 0.0138.$$

We report the largest discrepancy across bins and eligible thresholds. This is a finite-bin empirical approximation to α-local pivotality and is not an estimate of the theoretical $\sup_x$ quantity.

For this controlled experiment we also report a stronger full-CDF pivotality diagnostic. We evaluate a grid of queries varying scale, aspect ratio, and orientation separately, draw fresh residuals from the known conditional distribution at every grid point, and compute the largest KS distance between each conditional score distribution and the pooled score distribution. This diagnostic evaluates the whole score CDF and is therefore stronger than the fixed-miscoverage quantity used in the main experiments.

Nuisance-coordinate stress test. We repeat the experiment with $d_{\text{nuis}} \in \{0, 4, 8, 16, 32, 64\}$ while keeping the signal coordinates and residual draws paired across nuisance dimensions. The added coordinates never enter $\Sigma(x)$ and therefore contain no information about P_x.

Privileged retrieval remains essentially unchanged, with mean area approximately $1.06\times$ oracle throughout the sweep. Raw-X retrieval increases from $1.09\times$ at zero nuisance dimensions to $1.52\times$, $1.72\times$, $2.04\times$, $2.52\times$, and $3.20\times$ as the nuisance dimension increases. FacCP obtains

Table 3: Controlled experiment without nuisance coordinates. Area is normalized by the oracle geometry.

Method	α-local defect	Full-CDF defect	Mean area / oracle
Oracle geometry	0.020	0.017	1.02
Global geometry	0.139	0.361	3.95
Privileged retrieval	0.020	0.074	1.06
Raw-X retrieval	0.031	0.123	1.05
FacCP	0.029	0.117	1.07
Covariance prediction	0.022	0.044	1.04

$1.07\times$, $1.08\times$, $1.11\times$, $1.14\times$, $1.15\times$, and $1.19\times$, respectively. At 16 nuisance dimensions, the worst conditional-coverage deviation is 0.301 for raw-X retrieval, 0.066 for FacCP, and 0.035 for Privileged retrieval. Panels E–H of Fig. 2 show the corresponding area growth and representative prediction sets.

Localized-quantile comparison. We separately fix the global covariance and allow the calibration threshold to depend on the query using BaseLCP, calibrated LCP, and randomized LCP. Each method is evaluated using Privileged retrieval coordinates, raw X, and the learned FacCP representation. This comparison uses a common subset of $2{,}000$ calibration and $2{,}000$ evaluation observations.

The unlocalized fixed score has full-CDF pivotality defect 0.374. The localized procedures reduce the post-localization defect to between 0.130 and 0.199, depending on the localization method and representation. They nevertheless retain the same underlying family of globally oriented ellipses. To remove local-quantile estimation error from this comparison, we also compute the pointwise conditional 95% quantile of the fixed global score using fresh conditional draws. Its mean matched-coverage area remains $1.99\times$ oracle. Panels A–B of Fig. 2 show why. Local calibration can resize the fixed geometry and is effective for a large near-circular query, but cannot align the set with a tilted anisotropic residual distribution. Panels C–D show that methods which adapt the residual geometry can instead change its orientation and aspect ratio before the final conformal inflation.

C.2 Covariate-Shift Sweep

We report the full covariate-shift sweep from Section 5.1. Calibration is always performed on the source distribution $U \sim \mathrm{Unif}([-1,1]^3)$. At test time, one uncertainty coordinate U_j is replaced by the exponentially tilted distribution

$$p_\beta(u_j) = \frac{\exp(\beta u_j)}{\int_{-1}^{1} \exp(\beta v)\, dv}, \qquad u_j \in [-1,1],$$

with all remaining coordinates unchanged. Thus the experiment changes P_X while preserving $P_{R|X}$. The same source-calibrated conformal threshold is used for every target shift. Tables 4–6 report target coverage, averaged over three repetitions; each repetition contains $2{,}000$ target samples. Privileged retrieval localizes using the known uncertainty-generating coordinates, while Raw-X retrieval and covariance prediction use the same baselines as in Section 5.1.

C.3 Planar Quadcopter experiment details

Problem and retrieval. The planar quadcopter has a 6-dimensional state and 2-dimensional control. We use $H = 100$. The retrieval query consists of the nominal state rollout and control sequence, with

$$101 \cdot 6 + 100 \cdot 2 = 806$$

input coordinates. FacCP maps this query to an 8-dimensional retrieval representation. The residual contains the 6-dimensional state error over the 100 predicted steps, giving 600 residual coordinates. For every query, the same retrieval weights in Eq. (2) define a conditional empirical residual law. The geometry constructor is then changed without retraining the representation.

Table 4: Target coverage (%) under a shift toward larger residual scale. No target-domain recalibration is performed.

β	Global	FacCP	Raw-X	Privileged	Cov. Pred.
0.00	94.30	94.95	94.75	94.83	94.73
1.00	91.60	94.57	94.60	94.47	94.43
1.22	90.68	94.87	94.87	94.43	94.80
1.44	90.73	94.73	94.87	94.28	94.88
1.67	89.53	94.70	94.53	94.52	94.87
1.89	89.20	94.97	95.18	94.55	94.95
2.11	88.27	94.22	94.35	93.87	94.17
2.33	88.88	95.08	94.98	94.92	95.12
2.56	88.02	95.18	94.87	94.40	94.97
2.78	87.92	94.53	94.60	94.10	94.67
3.00	87.25	94.90	94.92	94.35	94.70

Table 5: Target coverage (%) under a shift toward larger residual aspect ratio. The conformal threshold remains fixed at its source-calibrated value.

β	Global	FacCP	Raw-X	Privileged	Cov. Pred.
0.00	95.12	94.78	94.77	95.02	94.87
1.00	92.12	95.35	95.58	94.73	94.48
1.22	92.33	95.55	95.77	95.30	94.93
1.44	91.17	95.17	95.42	94.55	94.33
1.67	91.12	95.50	95.85	95.03	94.42
1.89	91.27	95.75	96.02	95.27	94.82
2.11	90.35	95.65	95.80	94.58	93.90
2.33	90.42	95.97	96.15	95.05	94.33
2.56	89.97	96.10	96.22	94.98	94.67
2.78	89.27	96.08	96.30	95.12	94.77
3.00	89.22	96.38	96.33	95.27	94.60

Trajectory-level conformal calibration. For constructors producing a time-indexed raw score u_t, we use the single nonconformity score approach (Lindemann et al., 2025). A held-out portion of the calibration set determines positive time weights λ_t, after which each calibration trajectory has scalar score

$$S_i = \max_{1 \le t \le H} \lambda_t u_{it}.$$

The remaining calibration trajectories determine the split-conformal quantile C. The resulting trajectory is covered when

$$u_t \le \frac{C}{\lambda_t}, \qquad t = 1, \dots, H.$$

Thus one conformal event controls the complete trajectory and no union bound over time is used. For the geometry experiments we use 15% of calibration trajectories for the time normalization and the remaining 85% for the conformal quantile. Local and global methods use the same calibration procedure and differ only in whether the underlying residual geometry depends on the query.

Ellipsoids. At time t, retrieved residuals define a weighted center $c_t(x)$ and covariance

$$\Sigma_t(x) = \sum_j w_j(x) \big(r_{jt} - c_t(x)\big) \big(r_{jt} - c_t(x)\big)^\top + \epsilon I.$$

The raw score is the Mahalanobis radius

$$u_t(x, r_t) = \sqrt{\big(r_t - c_t(x)\big)^\top \Sigma_t(x)^{-1} \big(r_t - c_t(x)\big)}.$$

After trajectory-level calibration, the residual set is

$$G_t(x) = \left\{ r : \big(r - c_t(x)\big)^\top \Sigma_t(x)^{-1} \big(r - c_t(x)\big) \le \left(\frac{C}{\lambda_t}\right)^2 \right\}.$$

Table 6: Target coverage (%) under an orientation shift. The Global score is comparatively insensitive to this shift because its pooled geometry is approximately rotation-symmetric.

β	Global	FacCP	Raw-X	Privileged	Cov. Pred.
0.00	95.38	95.15	95.20	95.15	95.23
1.00	95.30	94.68	94.80	94.95	94.62
1.22	94.97	94.18	94.60	94.43	94.27
1.44	94.57	94.33	94.40	94.40	94.33
1.67	95.03	94.67	94.27	94.73	94.48
1.89	95.17	94.78	95.03	95.10	94.72
2.11	94.87	94.37	94.13	94.47	94.07
2.33	94.87	94.87	94.87	95.02	94.63
2.56	95.12	94.52	94.77	94.88	94.42
2.78	94.87	94.80	94.70	95.05	95.12
3.00	94.65	94.53	94.50	94.53	94.00

We use covariance ridge 10^{-3}. The global baseline applies the identical construction after pooling the geometry-bank residuals instead of retrieving them conditionally.

Hyperboxes. Retrieved residuals define a weighted center $c_t(x)$ and coordinatewise base half-widths $b_{tk}(x)$ from the 0.95 empirical quantile of the absolute residual deviations. The trajectory score is

$$S(x, r_{1:H}) = \max_{t,k} \frac{|r_{tk} - c_{tk}(x)|}{b_{tk}(x) + \epsilon}.$$

A single conformal quantile C therefore gives

$$G_t(x) = \left\{ r : |r_k - c_{tk}(x)| \le C\big(b_{tk}(x) + \epsilon\big),\ \forall k \right\}.$$

This score is already trajectory-level, so no additional time normalization is required.

KDE sets. At each time, the retrieved residuals form a weighted diagonal Gaussian KDE. We use the equivalent relative negative log-density score

$$u_t(x, r_t) = -\log \sum_j w_j(x) \exp\left(-\frac{1}{2} \left\| H_t^{-1}(r_t - r_{jt}) \right\|_2^2 \right),$$

where H_t is a diagonal Silverman bandwidth with a coordinatewise floor. Since the omitted Gaussian normalizer depends only on time, this score has the same timewise density sublevel sets as the KDE. SNSA calibration gives the nonconvex set

$$G_t(x) = \{ r : u_t(x, r) \le C/\lambda_t \}.$$

The global KDE uses uniform weights over a pooled residual bank, with at most 1024 components.

Zonotopes. Retrieved residuals define a weighted center $c_t(x)$ and a square PCA generator matrix $A_t(x)$. The base zonotope is

$$Z_t(x) = c_t(x) + A_t(x)[-1, 1]^6.$$

We score the multiplicative excess outside this base geometry by

$$u_t(x, r_t) = \left(\left\| A_t(x)^{-1}(r_t - c_t(x)) \right\|_\infty - 1 \right)_+.$$

SNSA calibration therefore produces

$$A_t(x) = c_t(x) + \left(1 + \frac{C}{\lambda_t} \right) A_t(x)[-1, 1]^6.$$

We use ridge 10^{-6} when constructing the generators.

Input-space localization baselines. To isolate the learned retrieval representation from the downstream geometry, we also construct ellipsoids by retrieving neighbors directly from the flattened 806-dimensional nominal state and control trajectory. We evaluate raw Euclidean distance and Euclidean distance after standardizing each state and control channel using geometry-bank statistics. All residual geometry construction and conformal calibration are otherwise unchanged. Raw and standardized input-space retrieval obtain mean log volumes -26.66 and -26.84, respectively, compared with -25.82 for global geometry and -33.25 for learned retrieval. Their held-out simultaneous coverages are 98.7%, 98.5%, 98.4%, and 99.1%, respectively.

Table 7: Planar Quadcopter capacity sweeps. Entries report FacCP / baseline. Coverage is simultaneous trajectory coverage in percent and $\log V$ is mean 6-dimensional log volume.

$N_\theta^{\mathrm{FacCP}}$	N_θ^{base}	$\log V$	Coverage
resCONTRA			
101	100	$-25.85/-17.96$	98.9/98.9
296	298	$-28.37/-17.19$	99.0/98.9
1,006	1,009	$-30.83/-16.53$	98.9/99.0
1,990	1,986	$-32.11/-20.70$	98.6/99.0
5,002	4,995	$-32.80/-24.07$	98.7/98.9
13,040	12,952	$-32.78/-25.75$	98.9/98.9
34,528	34,484	$-33.11/-27.28$	98.7/99.3
66,256	66,132	$-32.84/-27.55$	98.6/99.1
169,096	169,044	$-33.01/-28.34$	98.8/99.0
262,568	262,756	$-33.05/-28.30$	98.6/99.1
376,520	375,688	$-32.77/-29.23$	98.4/99.0
1,037,128	1,038,100	$-32.65/-31.12$	99.1/98.9
2,025,416	2,025,284	$-32.93/-33.10$	98.7/99.1
4,122,248	4,122,408	$-32.56/-32.91$	98.8/99.1
Covariance Prediction CP			
2,056	2,363	$-32.39/-26.86$	98.6/99.0
7,176	7,771	$-32.48/-27.95$	99.0/98.8
26,632	27,803	$-32.95/-28.90$	98.6/98.9
102,408	104,731	$-33.35/-29.55$	98.5/99.2
201,992	204,315	$-33.09/-31.68$	98.6/99.0
450,440	453,915	$-32.75/-32.23$	98.9/98.9
797,192	801,819	$-32.75/-33.47$	99.0/99.1
1,785,608	1,792,539	$-32.36/-33.43$	98.7/99.0
3,167,240	3,176,475	$-32.62/-33.85$	99.1/99.2
7,110,152	7,123,995	$-32.24/-33.37$	98.9/99.0

Capacity sweeps. We compare FacCP against resCONTRA (Fang et al., 2025) and Covariance Prediction CP (Srinivasan et al., 2026). resCONTRA learns a query-dependent flow transformation. For resCONTRA, we estimate the volume of the flow-defined conformal region by quasi-Monte Carlo integration of the inverse-flow Jacobian over the conformal latent region, rather than through a geometric approximation of the output set. Covariance Prediction CP directly predicts a query-dependent covariance parameterization and applies a Mahalanobis score. At each sweep point we choose architectures with closely matched learned parameter counts and report means over 10 training and evaluation repeats. The target trajectory coverage is 99%. Table 7 reports all completed capacity-sweep points.

Data sweeps. We fix capacity near the equal-volume crossover of each capacity sweep. For resCONTRA the matched models have approximately 2.03×10^6 parameters, while the Covariance Prediction CP comparison uses approximately 2.03×10^5. We vary the total data budget while holding validation, calibration, and test data fixed. Each FacCP budget includes a fixed 2,000-trajectory geometry bank, with the remaining trajectories used to train the retrieval representation. The baseline uses the corresponding total budget for direct model training. The data sweeps use 97.5% target trajectory coverage and 50% of the calibration trajectories for SNSA normalization. Results are averaged over 50 repeats.

Table 8: Planar Quadcopter data sweeps near the capacity crossover. Entries report FacCP / baseline. The training columns exclude FacCP's fixed 2,000-trajectory retrieval bank.

Total budget	Training trajectories	$\log V$	Coverage
	resCONTRA		
20%	700/2700	$-32.46/-27.11$	97.1/97.3
30%	2050/4050	$-32.04/-27.24$	97.1/97.7
40%	3400/5400	$-33.01/-29.21$	97.2/97.8
50%	4750/6750	$-32.70/-29.95$	97.5/97.8
60%	6100/8100	$-32.86/-30.40$	97.6/97.9
70%	7450/9450	$-32.54/-31.76$	97.4/98.1
80%	8800/10800	$-32.39/-32.34$	97.9/97.9
90%	10150/12150	$-32.44/-33.27$	97.6/97.9
100%	11500/13500	$-32.40/-33.40$	97.3/97.9
	Covariance Prediction CP		
20%	700/2700	$-32.84/-29.02$	97.0/97.5
30%	2050/4050	$-31.37/-29.67$	97.2/97.4
40%	3400/5400	$-32.92/-30.02$	97.1/97.5
50%	4750/6750	$-32.99/-30.21$	97.2/97.5
60%	6100/8100	$-32.85/-30.72$	97.2/97.8
70%	7450/9450	$-32.77/-31.24$	97.3/97.3
80%	8800/10800	$-32.91/-24.75$	97.3/97.0
90%	10150/12150	$-32.86/-31.83$	97.4/97.7
100%	11500/13500	$-32.73/-32.12$	97.4/97.8

α-local pivotality. At 99% target coverage, $\eta_n = 0.0341$. Over ten reference-fit difficulty bins, $\widehat{\Delta}_{\alpha,\eta_n}$ is 0.0942 for FacCP and 0.1570 for Global; the corresponding $\widehat{\Delta}_{\alpha,\eta_n} + \eta_n$ values are 0.1283 and 0.1911. At the fitted threshold, pooled coverage is $0.990/0.990$ (FacCP/Global), with maximum difficulty-bin coverage gaps $0.046/0.068$ and maximum smoothed difficulty gaps $0.023/0.010$. These diagnostics instantiate the empirical analogue of Cor. 10 above.

C.3.1 3D Quadcopter Experiment Details

Setup. The 3D quadcopter has a 12-dimensional state and 4-dimensional control with horizon $H = 200$. The nominal rollout and control sequence give

$$201 \cdot 12 + 200 \cdot 4 = 3212$$

query coordinates, while the residual trajectory contains $200 \cdot 12 = 2400$ coordinates. At each time step we construct a 12-dimensional ellipsoid and apply the same trajectory-level conformal calibration as in Appendix C.2, so a single conformal event controls the complete 200-step trajectory. We use at most 5,000 geometry trajectories, 2,000 calibration trajectories, and 1,000 test trajectories, with 35% of calibration trajectories used for time normalization and 99% target simultaneous coverage. Reported $\log V$ is the mean per-time-step 12-dimensional log volume.

Capacity sweeps. We match FacCP and baseline parameter counts as closely as possible. FacCP is already substantially more efficient than Mahalanobis CP at the smallest tested architecture (-74.50 vs. -62.15 at approximately 2.5×10^3 parameters). Mahalanobis improves with capacity but never closes the gap: even at approximately 1.6×10^6 parameters the mean log volumes are -78.19 and -74.67. resCONTRA behaves differently. Its gap contracts from -32.69 versus -76.02 at approximately 4.3×10^3 parameters to -78.14 versus -78.71 near 6×10^4, and it surpasses FacCP by approximately 9×10^4 parameters.

Data sweeps. We next fix capacity at approximately 1.6×10^6 parameters for the Mahalanobis comparison and 6×10^4 for resCONTRA, and vary the total training-data budget while keeping validation, calibration, and test sets fixed. FacCP remains more efficient than Mahalanobis throughout the sweep: at 20% of the data the mean log volumes are -73.16 versus -67.40, and at the full budget they are -78.02 versus -74.67. Against resCONTRA, FacCP has the low-data advantage

(-70.58 vs. -67.74 at 20%), while the gap closes as data increases; the methods are approximately matched near 70%, after which resCONTRA becomes slightly more efficient. Across both sweeps, simultaneous trajectory coverage remains near the 99% target.

α-local pivotality. At 99% target coverage, $\eta_n = 0.0280$. Over ten reference-fit difficulty bins, $\widehat{\Delta}_{\alpha,\eta_n}$ is 0.1099 for FacCP and 0.3166 for Global; the corresponding $\widehat{\Delta}_{\alpha,\eta_n} + \eta_n$ values are 0.1378 and 0.3446. At the fitted threshold, pooled coverage is $0.994/0.989$ (FacCP/Global), with maximum difficulty-bin coverage gaps $0.041/0.085$ and maximum smoothed difficulty gaps $0.049/0.013$. These diagnostics instantiate the empirical analogue of Cor. 10 above.

Table 9 reports the complete evaluated capacity sweep. For Mahalanobis CP, FacCP is already substantially more efficient at the smallest matched architecture: at approximately 2.5×10^3 parameters the mean log volumes are -74.50 and -62.15, respectively. Increasing Mahalanobis capacity narrows the gap, but does not eliminate it. Its best observed value is -76.09 at approximately 6.4×10^5 parameters, compared with -77.96 for FacCP at matched capacity. Even at approximately 1.6×10^6 learned parameters, Mahalanobis CP obtains -74.67 compared with -78.19 for FacCP. The persistence of the gap is therefore not explained simply by insufficient parameter count.

The resCONTRA sweep has qualitatively different behavior. At approximately 4.3×10^3 parameters, its mean log volume is -32.69 compared with -76.02 for FacCP, and at approximately 2.7×10^4 parameters the corresponding values are -71.97 and -78.29. The gap is nearly closed around 6×10^4 parameters, where resCONTRA reaches -78.14 compared with -78.71 for FacCP. With additional capacity, the ordering reverses: at approximately 9×10^4, 1.4×10^5, and 2.2×10^5 parameters, resCONTRA obtains mean log volumes of -81.44, -83.58, and -85.18, respectively, while FacCP remains near -78. Hence the two direct baselines fail for different reasons in this higher-dimensional setting: the Mahalanobis parameterization does not recover the same volume efficiency over the tested capacity range, whereas the more expressive flow eventually does, but only after a substantially larger low-capacity estimation burden.

Table 9: 3D quadcopter capacity sweeps. Entries report FacCP / baseline. Coverage is simultaneous coverage of the complete 200-step trajectory, and $\log V$ is mean per-time-step 12-dimensional log volume.

$N_\theta^{\mathrm{FacCP}}$	N_θ^{base}	$\log V$	Coverage (%)
		Mahalanobis CP	
2,514	2,506	$-74.50/-62.15$	99.1/98.9
69,972	69,961	$-77.74/-72.94$	99.2/99.4
100,002	100,004	$-77.39/-71.43$	99.1/99.0
160,008	159,974	$-77.83/-74.36$	99.2/99.1
250,013	249,982	$-78.39/-74.45$	99.2/99.2
400,006	399,991	$-77.53/-74.95$	99.1/99.0
639,997	639,994	$-77.96/-76.09$	99.1/99.3
1,000,006	1,000,004	$-78.45/-76.03$	99.0/99.2
1,600,023	1,599,991	$-78.19/-74.67$	99.1/98.9
		resCONTRA	
4,287	4,280	$-76.02/-32.69$	99.1/99.1
26,714	26,744	$-78.29/-71.97$	99.0/99.1
40,152	40,155	$-77.16/-72.07$	99.0/99.1
60,022	60,064	$-78.71/-78.14$	99.2/99.1
89,964	89,978	$-77.89/-81.44$	99.2/99.1
140,004	139,992	$-77.89/-83.58$	99.1/99.3
220,017	219,984	$-77.88/-85.18$	99.0/98.9

Table 10: 3D quadcopter data sweeps. Entries report FacCP / baseline. Coverage is simultaneous coverage of the complete 200-step trajectory, and $\log V$ is mean per-time-step 12-dimensional log volume.

Training budget (%)	Total episodes	$\log V$	Coverage (%)
Mahalanobis CP at $\sim 1.6M$ *parameters*			
20	5,400	$-73.16/-67.40$	99.3/99.1
30	8,100	$-73.87/-70.51$	99.2/99.1
40	10,800	$-77.00/-70.13$	99.2/99.2
50	13,500	$-77.48/-71.68$	99.2/98.8
60	16,200	$-77.58/-72.65$	99.1/99.0
70	18,900	$-77.91/-73.34$	99.2/99.1
80	21,600	$-77.22/-74.19$	99.2/99.0
90	24,300	$-77.38/-74.86$	99.2/99.4
100	27,000	$-78.02/-74.67$	99.1/98.9
resCONTRA at $\sim 60k$ *parameters*			
20	5,400	$-70.58/-67.74$	99.1/99.2
30	8,100	$-72.27/-71.61$	99.0/99.1
40	10,800	$-76.10/-72.13$	99.1/99.2
50	13,500	$-76.08/-73.71$	99.0/99.1
60	16,200	$-76.66/-75.79$	99.1/99.1
70	18,900	$-76.63/-76.86$	99.1/99.1
80	21,600	$-76.76/-78.24$	99.1/99.2
90	24,300	$-77.02/-77.51$	99.1/99.1
100	27,000	$-77.86/-78.39$	99.1/99.1

D AUTONOMOUS-DRIVING SUPPLEMENTARY RESULTS

The main comparison uses six candidates and fixed pricing. Here we give the construction behind those results, separate pricing from localization, and examine where the gains do and do not persist. The main comparison includes pricing for Global and Local. Here we show the corresponding unpriced ablation. MoE-CP retains its fixed $\eta = 1$ setting even where this parameter is omitted from plot labels.

D.1 DATASET AND ENCODER

The processed data contains 183,333 training and 22,979 validation scenarios. We train h_θ on the training split. We partition the validation split into 8,042 scenes reserved for encoder validation split and checkpoint selection, 6,894 reference scenes, 4,596 calibration scenes, and 3,447 internal TEST scenes. DEV uses separate subsets of the calibration pool: 900 scenes for conformal calibration and 1,000 for method evaluation. The final TEST evaluation uses all 3,447 reserved TEST scenes. All these partitions are disjoint at the scenario level.

The encoder takes focal history 21×18, neighbor histories $32 \times 21 \times 18$, map features $128 \times 20 \times 9$ and an anchor 60×2, with validity masks and map types. The 7.52M-parameter network uses width-256 tokens, four scene-attention layers and four trajectory-decoder layers, followed by a unit-normalized 16-dimensional embedding. We first pretrain the encoder for five epochs using masked reconstruction. For each scene, we retain 25% of the valid agent-history and map tokens as input and train the model to reconstruct the continuous features of the remaining masked tokens. Reconstruction uses a temporary two-layer decoder with hidden width 128, which is discarded after pretraining. After this, ten epochs of energy fine-tuning follow. The checkpoint is fixed across all targets.

For whitened residual trajectories $u_i \in \mathbb{R}^{120}$, fine-tuning uses

$$\mathcal{L}_i = \sum_j a_{ij} \frac{\|u_i - u_j\|_2}{\sqrt{120}} - \frac{1}{2} \sum_{j,\ell} a_{ij} a_{i\ell} \frac{\|u_j - u_\ell\|_2}{\sqrt{120}}.$$

Here, a_{ij} is the normalized weight assigned to reference scene j based on its embedding distance from query scene i. Each minibatch is split into two disjoint halves, with each half providing reference scenes for queries in the other. The reconstruction decoder is used only during pretraining and is then discarded. The subsequent objective trains the embeddings for residual-based retrieval; it does not directly minimize the footprint of the final conformal tubes.

D.2 Reference errors, geometry and SNSA

Global SNSA uses equally weighted reference scenes to estimate a covariance and a 95th-percentile Mahalanobis residual scale at each timestep. Local SNSA uses learned embedding weights to estimate both quantities separately for each query candidate. Global and Local SNSA use the same predicted paths as tube centers. They differ in how they weight reference errors to estimate each tube's shape and scale. To describe this construction, let $\mu_{j,k,t}$ denote the predicted position at time t for candidate k in reference scene j. Each scene contributes one residual path and one embedding:

$$e_{j,t} = y_{j,t} - \mu_{j,\kappa_j,t}, \qquad \bar{z}_j = h_\theta(x_j, \kappa_j), \qquad \kappa_j = \arg\min_k \frac{1}{60} \sum_t \|y_{j,t} - \mu_{j,k,t}\|_2.$$

The minimizing mode is fixed across the whole path. For the single-mode comparison, replace this choice with $\arg\max_k \pi_{j,k}$ and predict only the top-probability query mode. Neither procedure uses a test label when constructing its set. Query embeddings are $z_k(x) = h_\theta(x, k)$.

Covariance and weights. For a query mode k, Local assigns weights to all N bank scenes:

$$w_{k,j}(x) = \frac{\exp(-\|z_k(x) - \bar{z}_j\|_2^2/\tau)}{\sum_{\ell=1}^N \exp(-\|z_k(x) - \bar{z}_\ell\|_2^2/\tau)}, \qquad \tau = .30. \tag{25}$$

The same weights apply at all times, without top-k truncation or a uniform mixture. Write $\bar{e}_t = N^{-1} \sum_j e_{j,t}$ and $m_{k,t}(x) = \sum_j w_{k,j}(x) e_{j,t}$. The covariance estimates are

$$C_t^G = \frac{\sum_j (e_{j,t} - \bar{e}_t)(e_{j,t} - \bar{e}_t)^\top}{N - 1} + \varepsilon I_2, \tag{26}$$

$$C_{k,t}^L(x) = \frac{\sum_j w_{k,j}(x)(e_{j,t} - m_{k,t})(e_{j,t} - m_{k,t})^\top}{1 - \sum_j w_{k,j}(x)^2} + \varepsilon I_2, \qquad \varepsilon = .01\,\mathrm{m}^2. \tag{27}$$

Global estimates one covariance matrix for each of the 60 timesteps and uses these same matrices for every scene and candidate path. Local instead estimates a separate sequence of covariance matrices for each candidate path in each query scene. In both methods, reference errors are centered only to estimate covariance; the predicted MTR paths remain unchanged. Local uses the weighted covariance directly, without blending it with the Global estimate.

SNSA: normalizing residual scale over time. For either geometry, the query radius is $\rho_{k,t}(x, y) = \sqrt{(y_t - \mu_{k,t})^\top C_{k,t}(x)^{-1}(y_t - \mu_{k,t})}$. SNSA estimates a separate residual scale at each time:

$$r_{j,t}^G = \sqrt{e_{j,t}^\top (C_t^G)^{-1} e_{j,t}}, \qquad s_t^G = Q_{.95}(r_{1,t}^G, \ldots, r_{N,t}^G) + 10^{-6}, \tag{28}$$

$$r_{j,t}^L = \sqrt{e_{j,t}^\top (C_{j,t}^{L,-j})^{-1} e_{j,t}}, \qquad s_{k,t}^L(x) = Q_{.95}^{w_k(x)}(r_{1,t}^L, \ldots, r_{N,t}^L) + 10^{-6}. \tag{29}$$

Here $C_{j,t}^{L,-j}$ is bank scene j's local covariance after excluding its own residual and renormalizing the weights. Thus stored local radii use the reference scenes' own geometries; the query weights then select their weighted percentile. Global scales are shared across query scenes and modes; Local scales depend on the query mode. Q^w uses weighted interpolation, defined in the interpolation paragraph below. The normalizer quantile .95 estimates scale and consequently SNSA uses $\lambda = 1/s$.

Interpolation and bank draws. For weighted type-7 interpolation, sort radii $z_1 \le \cdots \le z_N$ with normalized positive weights ω_j, set $c_j = \sum_{\ell<j} \omega_\ell$ and $d = 1 - \omega_N$, and let $h = \min\{N-1, \max\{j : c_j \le pd\}\}$. Then

$$Q_p^w = z_h + \mathrm{clip}_{[0,1]}\left(\frac{pd - c_h}{\omega_h}\right)(z_{h+1} - z_h).$$

Uniform weights recover ordinary type-7 interpolation. Shrinking every scale by the same factor does not shrink the calibrated sets without pricing: at fixed geometry, $s \mapsto cs$ gives $\widehat{q} \mapsto \widehat{q}/c$. Efficiency therefore depends on relative scale allocation and covariance shape, not uniformly smaller scales.

For each paired draw, methods retain the same random 90% of bank scenes; the omitted 10% is unused.

D.3 CALIBRATION AND PRICING

With 900 calibration scenes and 60 timesteps, a Bonferroni correction yields unbounded finite-rank thresholds at targets $\geq 95\%$. Even when finite, simultaneous per-time coverage allows the covering candidate to change over time. Pricing adjusts each candidate's tube budget according to its predicted probability. We calibrate the resulting scene score on the separate calibration set. The score below covers both the priced results in the main table and the unpriced control, obtained by setting $\gamma = 0$.

Whole-path calibration. With $s_{k,t} = s_t^G$ or $s_{k,t}^L(x)$, $n = 900$ equally weighted calibration scenes give the score, threshold and prediction set, with pricing fee $b_k(x) = -\gamma \log \pi_k(x)$:

$$S_\gamma(x, y) = \min_{k \leq 6} \left[\max_{t \leq 60} \frac{\rho_{k,t}(x, y)}{s_{k,t}(x)} + b_k(x) \right], \qquad \widehat{q}_{\delta,\gamma} = S^{\text{cal}}_{\gamma,(\lceil (n+1)(1-\delta) \rceil)}, \tag{30}$$

$$\widehat{C}_{\delta,\gamma}(x) = \bigcup_{k:\, b_k(x) \leq \widehat{q}_{\delta,\gamma}} \{y : \rho_{k,t}(x, y) \leq (\widehat{q}_{\delta,\gamma} - b_k(x)) s_{k,t}(x)\ \forall t\}. \tag{31}$$

If the required calibration rank exceeds n, the threshold is $+\infty$. We compare the unpriced score ($\gamma = 0$) with fixed pricing ($\gamma = .10$), calibrating a separate threshold for each while keeping the encoder, covariance matrices and SNSA scales fixed. The scales determine the relative tube radii, while $\widehat{q}_{\delta,\gamma}$ sets their overall size to provide marginal whole-path coverage under exchangeability, with the scoring rule fixed before calibration. This guarantee applies across scenes, rather than separately within each scene group. Test labels are used only to evaluate coverage, not to construct the prediction sets.

Only the score and calibrated threshold change between $\gamma = 0$ and $\gamma = .10$. The same geometry, normalizers, candidates and calibration scenes are retained.

Why calibrate the complete path? Per-time coverage does not imply simultaneous coverage. A Bonferroni construction assigns miscoverage $\delta/60$ to each timestep and uses calibration rank $\lceil (n+1)(1-\delta/60) \rceil$. With $n = 900$, this rank exceeds 900 at the 95%, 97% and 99% targets, requiring an infinite threshold under the finite-rank convention. At 90% the rank is 900. Moreover, simultaneous membership in per-time unions still permits mode switching; it does not imply membership in one complete candidate tube. Our maximum-over-time, minimum-over-modes score calibrates the desired event directly. The reported native FORCE row does not apply Bonferroni: its low whole-path coverage reflects the difference in evaluation event, not a measured failure of a Bonferroni-corrected implementation.

D.4 BASELINES AND MEASUREMENT

FORCE-OPT (Chakraborty et al., 2026) is implemented through its offline allocation and per-time calibration core, without its online Bayesian monitor or planning system. Its whole-path coverage here is an additional evaluation event. MoE-CP (Kong et al., 2026) uses randomized gating-similarity calibration on the unpriced Global-SNSA score, with MTR probabilities as the gate. We retain the query infinity atom required by that construction. The reported $\eta = 1$ setting is exploratory: larger settings produced unbounded sets when weighted calibration support was insufficient. In our MoE-CP implementation, MTR candidate probabilities determine calibration-scene weights, while covariance and normalization scale remain shared across queries at each timestep.

The single-mode comparisons ask a further question: how do timestep tubes compare with sets that constrain an entire trajectory jointly? NLE (Messoudi et al., 2022) fits one joint 120D error ellipsoid using 5% nearest neighbors in standardized anchor/context features, blending 95% local with 5% pooled covariance; it retains empirical interpolated calibration. MVCS (Braun et al., 2026) learns a joint transform with frozen centers, bank whitening and finite-rank calibration.

We measure footprint as the spatial area occupied by the prediction set, integrated over time: $V(x) = 0.1 \sum_t \text{area}(\bigcup_k E_{k,t})$, in m^2 s. Here, $E_{k,t}$ denotes the ellipse for candidate k at timestep t, and overlapping regions are counted only once. For NLE and MVCS, footprint uses the two-dimensional

projections of the joint trajectory ellipsoid at each timestep, whereas coverage requires membership in the full joint set. Rank-6 coverage is evaluated on the 223 TEST scenes where the least-probable MTR candidate has the lowest average displacement error relative to the observed future. The single-mode evaluation uses this same scene subset, while predicting only the highest-probability path.

D.5 Does localization balance coverage across scenes?

A smaller average footprint does not by itself indicate more uniform coverage across scene types. We therefore examine both the similarity of score distributions near the target quantile and the empirical coverage within each scene bin.

Let $\widehat{F}_g$ denote the empirical score CDF in bin g, and $\widehat{F}$ the corresponding CDF over the full TEST cohort. We measure their largest discrepancy near the requested coverage level:

$$\widehat{D}_{\delta,\eta} = \max_{g\in\mathcal{G}} \sup_{z:\,|\widehat{F}(z)-(1-\delta)|\le\eta} \left|\widehat{F}_g(z) - \widehat{F}(z)\right|, \qquad \eta = 0.04749.$$

A smaller value indicates that bin-specific score distributions align more closely with the overall distribution near the target quantile. We include all 14 bins defined by intersection status, speed, predicted turn, neighboring-agent count and mode entropy.

Target	Global SNSA	Local SNSA (FacCP)
90%	12.40	**3.73**
95%	7.15	**2.83**
97%	7.04	**2.17**
99%	5.63	**1.69**

Table 11: Quantile-local score discrepancy $\widehat{D}_{\delta,\eta}$ in percentage points. Therefore, lower is better.

Local SNSA has a smaller discrepancy at every target (Table 11), indicating more consistent score distributions across the evaluated scene types near the relevant quantiles. These results isolate the unpriced construction; they do not describe the priced scores used in the main table. We separately measure worst-bin coverage error, $\max_g |\widehat{p}_g - (1-\delta)|$, where $\widehat{p}_g$ is the empirical whole-path coverage in bin g at the calibrated threshold. This measures departure from nominal coverage, whereas $\widehat{D}_{\delta,\eta}$ measures differences between score distributions.

Coverage across scene bins. Table 12 reports coverage in all 14 TEST bins at the 95% target. Local improves intersection coverage, although six-mode coverage for fast-moving agents is slightly lower.

D.6 Single-mode comparison

Table 14 uses only the top-probability anchor for both reference errors and predictions. It compares the tube construction with joint-trajectory NLE/MVCS baselines without an advantage from extra modes.

Why is the raw-head MTR footprint smaller? All three single-mode methods use the same highest-probability predicted path, but construct different uncertainty regions around it. Raw-head MTR uses its exported covariance and a Gaussian waypoint radius, without calibrating coverage of the complete trajectory. Global and Local CP instead estimate covariance and timestep scales from observed reference errors and calibrate a threshold for simultaneous coverage of all 60 steps. Their larger footprints therefore account for errors that the narrow raw-head ellipses often exclude. At the 95% target, raw-head MTR has a footprint of 26.4 m^2 s and only 1.68% path coverage, compared with 2,362.7/1,713.0 m^2 s and 95.76/96.09% for Global/Local CP. The smaller raw-head footprint is thus accompanied by severe undercoverage, not comparable protection at lower cost. This also omits the covariance clipping used in training, producing very narrow early ellipses.

		Single mode		Six modes	
Scene bin	Scenes	Global	Local (FacCP)	Global	Local (FacCP)
Non-intersection	2468	96.83	96.02	95.82	94.62
Intersection	979	93.05	96.26	93.16	94.95
Speed ≤ 5 m/s	1438	95.56	95.01	95.61	94.55
Speed $(5, 12]$ m/s	1513	95.93	96.98	94.50	94.85
Speed > 12 m/s	496	95.78	96.48	95.20	94.78
Turn $\leq 15^\circ$	2843	96.93	96.06	95.86	94.61
Turn $(15, 45]^\circ$	272	94.32	94.37	93.34	94.85
Turn $> 45^\circ$	332	86.90	97.67	89.68	95.50
Neighbors 0–2	1457	93.80	95.18	93.29	94.33
Neighbors 3–6	1279	96.54	96.75	95.89	94.56
Neighbors ≥ 7	711	98.36	96.76	97.21	95.78
Entropy low	1184	97.14	96.83	96.54	96.00
Entropy middle	1074	95.65	95.58	94.94	93.87
Entropy high	1189	94.46	95.80	93.71	94.20

Table 12: Whole-path coverage (%) at the 95% target on internal TEST, averaged over 16 reference-bank draws. Global denotes Global SNSA; Local (FacCP) denotes Local SNSA.

Target	Method	γ	Footprint $\pm$ SD	Path (%)	Rank-6 (%)
90	Global SNSA	0.00	493.2 $\pm$ 10.2	89.86	69.65
90	Global SNSA	0.10	419.9 $\pm$ 9.3	89.51	64.38
90	Local SNSA (FacCP)	0.00	487.4 $\pm$ 5.4	89.50	73.82
90	Local SNSA (FacCP)	0.10	399.6 $\pm$ 6.2	89.32	69.03
95	Global SNSA	0.00	749.5 $\pm$ 14.3	95.06	86.04
95	Global SNSA	0.10	683.8 $\pm$ 9.6	95.19	84.16
95	Local SNSA (FacCP)	0.00	697.9 $\pm$ 11.1	94.72	83.83
95	Local SNSA (FacCP)	0.10	581.9 $\pm$ 11.3	94.46	81.39
97	Global SNSA	0.00	985.5 $\pm$ 16.1	96.98	90.75
97	Global SNSA	0.10	880.0 $\pm$ 15.1	97.02	90.02
97	Local SNSA (FacCP)	0.00	1,003.1 $\pm$ 10.3	97.28	92.35
97	Local SNSA (FacCP)	0.10	870.0 $\pm$ 17.9	97.27	90.89
99	Global SNSA	0.00	1,795.5 $\pm$ 57.1	99.19	97.06
99	Global SNSA	0.10	1,651.1 $\pm$ 54.9	99.21	96.92
99	Local SNSA (FacCP)	0.00	1,608.1 $\pm$ 21.9	99.25	97.23
99	Local SNSA (FacCP)	0.10	1,416.4 $\pm$ 19.5	99.24	97.28

Table 13: Six-mode pricing comparison on internal TEST. Footprint is in m^2 s.

Target	Method	Footprint $\pm$ SD	Path (%)	Rank-6 (%)
90	MTR raw-head	20.3	1.19	0.00
90	Global SNSA	1,277.2 $\pm$ 24.2	89.98	63.03
90	Local SNSA (FacCP)	1,010.9 $\pm$ 18.9	91.64	66.84
90	Messoudi NLE	27,369.2 $\pm$ 828.8	90.71	86.24
90	MVCS-L2	84,668.7 $\pm$ 20,470.2	91.05	84.87
95	MTR raw-head	26.4	1.68	0.00
95	Global SNSA	2,362.7 $\pm$ 34.9	95.76	80.77
95	Local SNSA (FacCP)	1,713.0 $\pm$ 29.7	96.09	84.78
95	Messoudi NLE	47,625.8 $\pm$ 1,276.1	95.21	92.74
95	MVCS-L2	168,608.7 $\pm$ 41,988.4	96.39	95.57
97	MTR raw-head	30.9	2.15	0.00
97	Global SNSA	3,356.1 $\pm$ 72.4	97.63	88.31
97	Local SNSA (FacCP)	2,310.1 $\pm$ 36.2	97.53	89.41
97	Messoudi NLE	81,456.4 $\pm$ 3,062.1	97.78	96.33
97	MVCS-L2	231,938.9 $\pm$ 62,693.7	98.08	97.98
99	MTR raw-head	40.6	3.25	0.00
99	Global SNSA	5,952.4 $\pm$ 123.2	99.06	95.57
99	Local SNSA (FacCP)	4,294.0 $\pm$ 116.5	99.19	96.08
99	Messoudi NLE	150,365.8 $\pm$ 6,536.7	99.63	99.27
99	MVCS-L2	368,926.5 $\pm$ 107,468.6	99.40	99.22

Table 14: Single-mode internal TEST comparison, without pricing. Footprint is in m^2 s.

D.7 Why joint ellipsoids have large footprints?

Figure 7 illustrates why spatial footprint and joint trajectory volume can rank methods differently. Footprint measures the area available at each timestep and sums these areas over time; it does not capture dependencies between positions at different timesteps. Joint volume instead measures the set of complete trajectories allowed by those dependencies.

For Global and Local SNSA, the single-mode set is a product of 60 timestep ellipses: any sequence of positions lying inside the respective ellipses belongs to the set. NLE and MVCS instead use a joint ellipsoid that couples all 120 trajectory coordinates. Its projection at a given timestep can be large, even though positions chosen independently from different projections may not form an admissible trajectory. A thin, tilted ellipse provides a simple analogy: it can have wide projections onto both axes while enclosing little area. Cross-time dependence produces the same effect in the joint trajectory ellipsoid.

In figure 7, all four methods cover the observed trajectory. Global and Local footprints are 3,338.0 and 776.6 m^2 s, compared with 29,011.9 and 52,215.6 for NLE and MVCS. Their joint log-volumes, $\log(V_{120}/\mathrm{m}^{120})$, are respectively 297.61, 222.16, -182.06 and -117.25. Thus, NLE and MVCS allow a wider spatial region at individual timesteps but a smaller volume of complete trajectories.

Neither metric determines the other: footprint describes spatial occupancy over time, whereas joint volume also reflects restrictions on how positions can vary together.

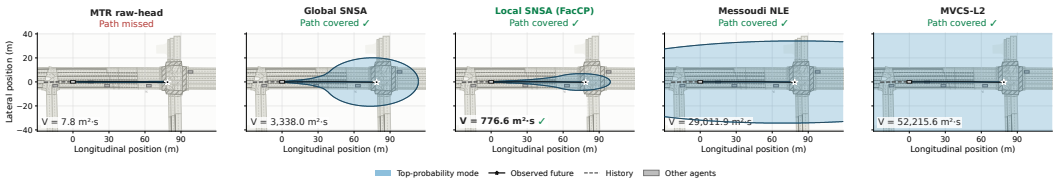


Figure 7: Single-mode example at the 97% target. Local covers the whole path with 76.7% less footprint than Global. NLE and MVCS show spatial projections of joint trajectory ellipsoids.

D.8 What changes when pricing is added?

Pricing uses the forecaster's candidate probabilities to allocate tube radii. We add $b_k = -\gamma \log \pi_k$ to each candidate's score, so a high-probability candidate incurs a smaller penalty and retains more of the calibrated radius budget. A low-probability candidate incurs a larger penalty and receives a smaller budget, potentially becoming inactive. This favors candidates that the forecaster regards as more plausible, reducing the footprint spent on unlikely alternatives. The allocation changes uncertainty around the predicted paths; it does not move the paths themselves.

We calibrate the resulting scene score on a separate calibration set, preserving the marginal whole-path coverage guarantee under exchangeability. This guarantee concerns coverage by at least one candidate tube across scenes; it does not require equal protection for every candidate or scene group. Pricing therefore offers a probability-informed allocation, rather than a guarantee of balanced coverage across high- and low-probability outcomes.

Table 13 isolates this effect by holding the encoder, covariance matrices and SNSA scales fixed and comparing $\gamma = 0$ with $\gamma = .10$. At the 95% target, pricing reduces Local's mean footprint from 697.9 to 581.9 m^2 s, while empirical whole-path coverage changes from 94.72% to 94.46%. Rank-6 coverage falls from 83.83% to 81.39% on the 223 scenes whose best-matching candidate is least probable. This illustrates both the efficiency benefit and the reduced protection for outcomes assigned low probability.

Both settings use the same 900 calibration scenes, all 3,447 TEST scenes and 16 paired reference-bank subsets, with a separate threshold calibrated for each setting, draw and target. Pricing strength is fixed before TEST evaluation. Because calibration also changes the overall threshold, pricing need not shrink every individual tube.