Robust Conditional Conformal Prediction via Branched Normalizing Flow

Rui Xu[1] **Xingyuan Chen**[1] **Wenxing Huang**[2]
Minxuan Huang[2] **Weiyan Chen**[2] **Sihong Xie**[1] **Hui Xiong**[1]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Guangzhou Medical University

Abstract

Conformal prediction (CP) constructs prediction sets with marginal coverage guarantees under exchangeability of calibration and test samples. Under distribution shift, existing alignment approaches primarily target marginal conformal score distributions, leaving conditional coverage at individual test inputs uncontrolled. As a consequence, CP can remain unreliable in regions where the conditional score distributions are mismatched. In this work, we bound the conditional invalidity of CP under distribution shift in terms of the Wasserstein distance between the calibration and test distributions. This result highlights the role of invertible transport in mitigating conditional coverage degradation. Motivated by this insight, we introduce Branched Normalizing Flow (BNF), a two-branch architecture that normalizes a test input to the calibration distribution and transforms the prediction set of the normalized input back to the test distribution while preserving an oracle conditional guarantee under exact population transport. Empirically, BNF consistently improves conditional coverage robustness on nine datasets across a wide range of confidence levels.

1 Introduction

Due to data noise and lack of prior knowledge, prediction uncertainty hinders applications of AI in various safety-critical domains. Conformal Prediction (CP) yields a set of possible targets rather than a single prediction to accommodate prediction uncertainty [1, 2]. We focus on CP for **regression** [3]. Given a trained model h, a score function $s(X, Y) = |h(X) - Y|$ computes the residuals (conformal scores) of n calibration instances $\{(X_i, Y_i)\}_{i=1}^n$. Denoting τ the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the conformal scores, a vanilla prediction set $\mathcal{C}_{\mathrm{M}}(X_{n+1})$ of a test input X_{n+1} contains all target values whose conformal scores are smaller than τ. Let P_{XY} and Q_{XY} be calibration and test distributions in space $\mathcal{X} \times \mathcal{Y}$, respectively. If the data are independent and identically distributed (i.i.d.) so that $P_{XY} = Q_{XY}$, the prediction set $\mathcal{C}_{\mathrm{M}}(X_{n+1})$ achieves the **marginal coverage guarantee** $\Pr\left(Y_{n+1} \in \mathcal{C}_{\mathrm{M}}(X_{n+1})\right) \geq 1 - \alpha$. However, since τ does not depend on the specific test input x, $\mathcal{C}_{\mathrm{M}}(X_{n+1})$ has constant size and lacks adaptiveness. To address the weakness, conditional prediction set $\mathcal{C}_{\mathrm{C}}(X_{n+1})$ aims at **conditional coverage guarantee** $\Pr(Y_{n+1} \in \mathcal{C}_{\mathrm{C}}(X_{n+1}) | \, X_{n+1} = x \,) \geq 1 - \alpha, \forall x \in \mathcal{X}$, which provides more effective uncertainty quantification [4, 5].

In practice, a distribution shift ($P_{XY} \neq Q_{XY}$) can violate the i.i.d. assumption. For example, **multi-source domain generalization (MSDG)** considers Q_{XY} as a random mixture of multiple source distributions [6]. In this scenario, ensuring coverage guarantees becomes both important and challenging. Let P_V and Q_V be the calibration and test conformal score distributions in space $\mathcal{V}$, respectively. The difference between the cumulative probabilities of P_V and Q_V at τ can measure the validity of marginal coverage. Various upper bounds of the discrepancy between P_V and Q_V are proposed to estimate the potential deviation from the nominal marginal coverage [7, 8]. Nevertheless,

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

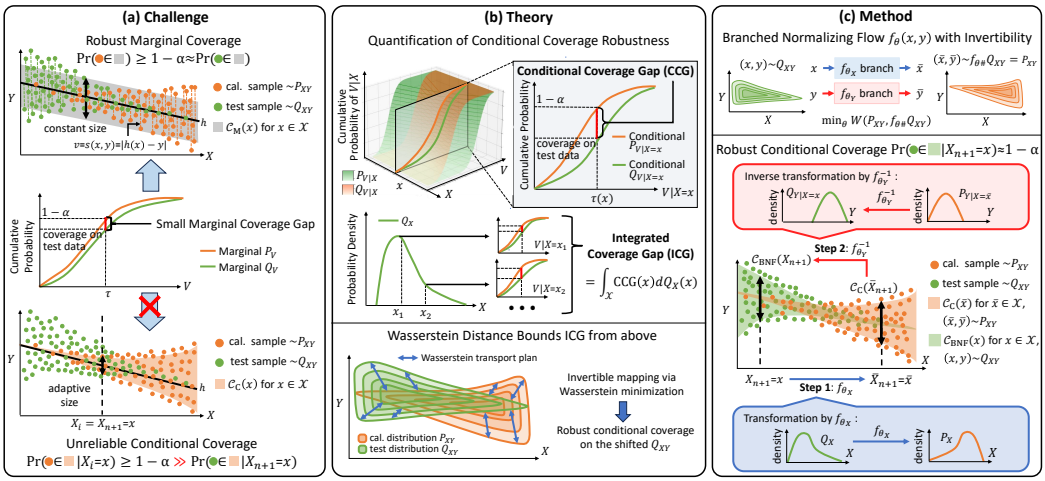


Figure 1: **(a)** Vanilla prediction set $\mathcal{C}_{\mathrm{M}}(x)$ has constant size and offers marginal coverage, which is robust if conformal score distributions P_V and Q_V have similar cumulative probabilities at τ. The oracle conditional prediction set $\mathcal{C}_{\mathrm{C}}(x)$ has input-dependent size and provides conditional coverage for calibration inputs $X_i = x$ where $i = 1, ..., n$, but may fail on non-i.i.d. test input $X_{n+1} = x$. The difference between P_V and Q_V cannot capture the reliability of conditional coverage on the shifted test data; **(b)** Conditional coverage gap (CCG) measures $\mathcal{C}_{\mathrm{C}}(X_{n+1})$ validity at x by comparing $P_{V|X=x}$ and $Q_{V|X=x}$. Integrated coverage gap (ICG) is the expectation of CCG under Q_X for a holistic robustness measure. Wasserstein distance $W(P_{XY}, Q_{XY})$ bounds ICG to reveal how a distribution shift results in non-i.i.d. conformal scores. An invertible mapping between P_{XY} and Q_{XY} via Wasserstein minimization promotes robust conditional coverage; **(c)** Branched Normalizing Flow f_θ minimizes $W(f_{\theta\#}Q_{XY}, P_{XY})$, where $f_{\theta\#}Q_{XY}$ is a pushforward distribution. Assuming exact joint alignment, then during inference, we first compute a normalized test input $\overline{X}_{n+1}$ by f_{θ_X} and generate $\mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1}) \subseteq \mathcal{Y}$ with conditional guarantee on P_{XY}. Then, $f_{\theta_Y}^{-1}$ inversely transforms the set to $\mathcal{C}_{\mathrm{BNF}}(X_{n+1}) \subseteq \mathcal{Y}$ with preserved conditional coverage on Q_{XY}.

since existing methods align calibration and test conformal scores solely at the marginal level, they do not elucidate how score distributions change with individual inputs, leaving conditional score behavior largely unaddressed. As a result, these methods are unable to assess the conditional coverage of $\mathcal{C}_{\mathrm{C}}(X_{n+1})$ under distribution shift (Figure 1(a)).

We study conditional coverage robustness under distribution shifts through three contributions.

1. **Quantification of conditional coverage robustness.** We define the conditional coverage gap (CCG) to measure conditional invalidity under distribution shift, and the integrated coverage gap (ICG) as the expectation of CCG under the test distribution. (Figure 1(b), 1st plot).
2. **Bounding by Wasserstein distance.** We bound ICG with the Wasserstein distance between calibration and test distributions to reveal how a distribution shift is propagated from data space to score space. This bound implies that an invertible mapping can promote robust conditional coverage (Figure 1(b) 2nd plot).
3. **Branched Normalizing Flow (BNF).** We embed the Wasserstein bound into a branched structure, defined as $f_\theta(x, y) = (f_{\theta_X}(x), f_{\theta_Y}(y)) = (\overline{x}, \overline{y})$, to transform Q_{XY} to P_{XY} (Figure 1(c) 1st plot). The structure does not explicitly couple the transformations of x and y, so f_{θ_X} can compute the normalized test input without knowing the true label during inference. If the conditional prediction set of the normalized input holds $1 - \alpha$ conditional coverage on the calibration distribution, $f_{\theta_Y}^{-1}$ inherits this guarantee on the test distribution under exact joint transport (Figure 1(c) 2nd plot).

Experiments on nine datasets cover both synthetic distribution shifts [9] and real-world challenges, including sales prediction across time series [10], traffic forecasting with mismatched data [11], medicine decision-making for different populations [12, 13], and epidemic modeling over pandemic intervals [14], demonstrating improved conditional guarantee robustness under distribution shift.

2 Background

2.1 Conditional Conformal Prediction

Denote $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ the input and output random variables, respectively. With a regression model $h : \mathcal{X} \to \mathcal{Y}$, a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathcal{V} \subseteq \mathbb{R}$ outputs conformal scores to assess how data conform to the model h. We denote $V \in \mathcal{V}$ the random variable of conformal score, typically defined as the absolute residual: $V := s(X, Y) = |h(X) - Y|$. With instances $\{(X_i, Y_i)\}_{i=1}^n$ from a calibration distribution P_{XY}, split conformal prediction computes calibration conformal scores $V_i = s(X_i, Y_i)$ for $i = 1, ..., n$ [15]. For a test instance (X_{n+1}, Y_{n+1}), a vanilla prediction set is given by $\mathcal{C}_{\mathrm{M}}(X_{n+1}) := \{y : s(X_{n+1}, y) \leq \tau, y \in \mathcal{Y}\}$, where τ is the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of $\{V_i\}_{i=1}^n$.[1] Under exchangeability of calibration and test samples, including the i.i.d. setting considered here where $(X_i, Y_i) \sim P_{XY}$ for $i = 1, \ldots, n+1$, $\mathcal{C}_M(X_{n+1})$ provides a **marginal coverage guarantee** for the ground truth Y_{n+1}, namely,

$$\Pr\left(Y_{n+1} \in \mathcal{C}_{\mathrm{M}}(X_{n+1})\right) = \Pr(V_{n+1} \leq \tau) \geq 1 - \alpha. \tag{1}$$

As τ is independent from test inputs, fixed-size prediction sets often underestimate uncertainty for hard samples and overestimate it for easy ones [16]. Therefore, a conditional prediction set $\mathcal{C}_{\mathrm{C}}(X_{n+1})$ aims at improving the guarantee under the condition where $X_{n+1} = x$, $\forall x \in \mathcal{X}$. Theoretically, denote $\tau(x)$ the $(1-\alpha)$ quantile of $P_{V|X=x}$. Then, for $X_{n+1} = x$, a conditional prediction set is

$$\mathcal{C}_{\mathrm{C}}(X_{n+1}) := \{y : s(X_{n+1}, y) \leq \tau(x), y \in \mathcal{Y}\}, \tag{2}$$

with the **conditional guarantee** under i.i.d. assumption:

$$\Pr\left(Y_{n+1} \in \mathcal{C}_{\mathrm{C}}(X_{n+1}) | X_{n+1} = x\right) = \Pr\left(V_{n+1} \leq \tau(x) | X_{n+1} = x\right) \geq 1 - \alpha, \forall x \in \mathcal{X}. \tag{3}$$

Eq. (3) describes a population-level conditional guarantee based on the true conditional quantile. With finite calibration samples, nontrivial distribution-free pointwise conditional coverage is generally unattainable. Regularity assumptions, such as Lipschitz continuity of the conditional density of Y given $X = x$, can support approximations to this guarantee [5, 17, 18].

2.2 Conformal Prediction in Multi-Source Domains

Multi-source domain generalization (MSDG) is a case of joint distribution shifts, where $(X_{n+1}, Y_{n+1}) \sim Q_{XY} \neq P_{XY}$. Under MSDG, the test distribution is a random mixture of multiple sources, so standard conformal prediction cannot maintain its coverage guarantees without accounting for the shift. To address marginal coverage under such shifts, conservative CP approaches consider the worst-case scenario [19, 20]. Recent work further regularizes the model h for a balance between marginal coverage and prediction set size [8]. A related area is federated CP [21, 22], where robust CP is pursued across separated sources without data centralization. Nevertheless, the impact of joint distribution shifts on conditional coverage guarantees remains poorly understood. We therefore develop a theoretical framework to analyze conditional coverage robustness under such shifts.

3 Theory

3.1 Conditional Coverage Lower Bound

Let $P_{V|x}$ and $Q_{V|x}$ be the calibration and test conformal score distributions conditioned on an input x. Denote $F^P_{V|x}(\cdot)$ and $F^Q_{V|x}(\cdot)$ cumulative distribution functions (CDFs) of $P_{V|x}$ and $Q_{V|x}$, respectively. Using this notation, the conditional guarantee under i.i.d. assumption in Eq. (3) is reformulated by

$$F^P_{V|x}(\tau(x)) \geq 1 - \alpha. \tag{4}$$

To quantify how a distribution shift $(X_{n+1}, Y_{n+1}) \sim Q_{XY} \neq P_{XY}$ impedes the conditional guarantee at $X_{n+1} = x$, we define **conditional coverage gap (CCG)** by

$$\mathrm{CCG}(P, Q, x) := \left| F^P_{V|x}(\tau(x)) - F^Q_{V|x}(\tau(x)) \right|, \tag{5}$$

[1] Equivalently, τ can be defined as the $1-\alpha$ quantile of $\{V_i\}_{i=1}^n \cup \{V_\infty\}$ [1, 3].

which implies $F^Q_{V|x}(\tau(x)) \geq F^P_{V|x}(\tau(x)) - \text{CCG}(P, Q, x)$. Combining with Eq. (4), we obtain a lower bound on the conditional coverage for $Y_{n+1} \sim Q_{Y|x}$ by

$$\begin{aligned} &\Pr\left(Y_{n+1} \in \mathcal{C}_{\text{C}}(X_{n+1}) | X_{n+1} = x\right) = F^Q_{V|x}(\tau(x)) \\ &\geq F^P_{V|x}(\tau(x)) - \text{CCG}(P, Q, x) \geq 1 - \alpha - \text{CCG}(P, Q, x). \end{aligned} \tag{6}$$

Since test inputs follow Q_X, evaluating CCG at a single x cannot take $Q_X(x)$ at different x into account. Hence, **integrated coverage gap (ICG)** is defined as the expectation of CCG under Q_X by

$$\text{ICG}(P, Q) := \int_{\mathcal{X}} \text{CCG}(P, Q, x) \mathrm{d}Q_X(x). \tag{7}$$

By integrating CCG over Q_X, ICG is a comprehensive metric for the validity of $\mathcal{C}_{\text{C}}(X_{n+1})$. A low ICG means that conditional coverage is consistently close to $1 - \alpha$ over Q_X.

3.2 Linking Data Shift to Coverage via Wasserstein

We further explore how a distribution shift between P_{XY} and Q_{XY} in space $\mathcal{X} \times \mathcal{Y}$ is propagated to a shift between $P_{V|x}$ and $Q_{V|x}$ in space $\mathcal{V}$ for all $x \in \mathcal{X}$.

Definition 3.1 (p-Wasserstein Distance between Population Distributions [23])**.** For any probability measures μ_X and ν_X defined on a metric space $(\mathcal{X}, d_{\mathcal{X}})$, where $\mathcal{X}$ is a set and $d_{\mathcal{X}}$ is a metric on $\mathcal{X}$, the Wasserstein distance of order $p \geq 1$ between μ_X and ν_X is defined by

$$W_p(\mu_X, \nu_X) = \inf_{\gamma \in \Gamma(\mu_X, \nu_X)} \left(\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_1, x_2)^p \, \mathrm{d}\gamma(x_1, x_2) \right)^{\frac{1}{p}},$$

where $\Gamma(\mu_X, \nu_X)$ is the set of all joint probability measures γ on $\mathcal{X} \times \mathcal{X}$ with marginals $\gamma(\mathcal{A} \times \mathcal{X}) = \mu_X(\mathcal{A})$ and $\gamma(\mathcal{X} \times \mathcal{B}) = \nu_X(\mathcal{B}), \forall$ measurable sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{X}$.

The Wasserstein distance with $p = 1$ is denoted as W. An upper bound of the marginal coverage gap is proposed in [8]. Let L be the Lebesgue density bound of $P_{V|x}$ for all $x \in \mathcal{X}$ [24]. We derive

$$\text{CCG}(P, Q, x) \leq \sqrt{2L \cdot W(P_{V|x}, Q_{V|x})}. \tag{8}$$

Next, we explore how $W(P_{V|x}, Q_{V|x})$ arises from the difference in $P_{Y|x}$ and $Q_{Y|x}$ by Theorem 3.2.

Theorem 3.2. *Let μ_{XY} and ν_{XY} be probability measures in the metric space $(\mathcal{X} \times \mathcal{Y}, d_{\mathcal{XY}})$, where $d_{\mathcal{XY}}$ is the 2-product metric of $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ such that $d_{\mathcal{XY}}((x_1, y_1), (x_2, y_2)) := ||(d_{\mathcal{X}}(x_1, x_2), d_{\mathcal{Y}}(y_1, y_2))||_2$. Let $s : \mathcal{X} \times \mathcal{Y} \to \mathcal{V}$ be a measurable function such that $s(x, y) = v$. In the metric space $(\mathcal{V}, d_{\mathcal{V}})$, denote μ_V the probability measure of $s(X, Y)$ for $(X, Y) \sim \mu_{XY}$. Also, let ν_V be the probability measure of $s(X, Y)$ for $(X, Y) \sim \nu_{XY}$. If s has a continuity constant κ at x such that $\frac{d_{\mathcal{V}}(s(x,y_1),s(x,y_2))}{d_{\mathcal{Y}}(y_1,y_2)} \leq \kappa, \forall x \in \mathcal{X}$ and $\forall y_1, y_2 \in \mathcal{Y}$, the following inequality holds:*

$$W(\mu_{V|x}, \nu_{V|x}) \leq \kappa \cdot W(\mu_{Y|x}, \nu_{Y|x}). \tag{9}$$

A related theorem in [8] does not condition on a specific x. Since $\mathcal{V}, \mathcal{Y} \subseteq \mathbb{R}$, we can take the metrics $d_{\mathcal{V}}(\cdot, \cdot)$ and $d_{\mathcal{Y}}(\cdot, \cdot)$ as the absolute value of the difference. Therefore, according to Theorem 3.2, if the score function $s(X, Y)$ is continuous with a constant κ such that $\frac{|s(x,y_1)-s(x,y_2)|}{|y_1-y_2|} \leq \kappa, \forall x \in \mathcal{X}, \forall y_1, y_2 \in \mathcal{Y}$, we derive that

$$W(P_{V|x}, Q_{V|x}) \leq \kappa \cdot W(P_{Y|x}, Q_{Y|x}). \tag{10}$$

For an intuitive explanation, a smaller κ implies that the score function s becomes less responsive to changes in y conditioned on x. Consequently, a substantial distribution shift between $P_{Y|x}$ and $Q_{Y|x}$ will not result in a large $W(P_{V|x}, Q_{V|x})$. Combining Eq. (10) and Eq. (8) leads to

$$\text{CCG}(P, Q, x) \leq \sqrt{2\kappa L \cdot W(P_{Y|x}, Q_{Y|x})}. \tag{11}$$

Besides, as $\sqrt{W(P_{Y|x}, Q_{Y|x})} \leq W(P_{Y|x}, Q_{Y|x}) + 1/4$,[2]. We can bound ICG based on Eq. (11) by

$$\text{ICG}(P, Q) \leq \sqrt{2\kappa L} \left(\int_{\mathcal{X}} W(P_{Y|x}, Q_{Y|x}) \mathrm{d}Q_X(x) + \frac{1}{4} \right),$$

[2]When $W(P_{Y|x}, Q_{Y|x}) \geq 1$, we obtain $\text{ICG}(P, Q) \leq \sqrt{2\kappa L} \int_{\mathcal{X}} W(P_{Y|x}, Q_{Y|x}) \mathrm{d}Q_X(x)$ by $\sqrt{W(P_{Y|x}, Q_{Y|x})} \leq W(P_{Y|x}, Q_{Y|x})$. This tightens Eq. (13) to $\text{ICG}(P, Q) \leq \sqrt{2\kappa L} \cdot \eta \cdot W(P_{XY}, Q_{XY})$

showing that transporting $P_{Y|x}$ to $Q_{Y|x}$ for all $x \in \mathcal{X}$ at population level is sufficient to eliminate conditional gap.

Estimating the conditional distributions $P_{Y|x}$ and $Q_{Y|x}$ from finite samples is challenging, particularly in moderate or high dimensions. To address this difficulty, we establish a bound through the joint Wasserstein distance under explicit conditional-coupling and diagonal-overlap assumptions on an optimal transport plan.

Theorem 3.3. *Let μ_{XY} and ν_{XY} be probability measures with finite first moments on $(\mathcal{X} \times \mathcal{Y}, d_{\mathcal{XY}})$, equipped with the 2-product metric. Denote their conditional label distributions by $\mu_{Y|x}$ and $\nu_{Y|x}$. Let $\gamma^*_{XYXY} \in \Gamma(\mu_{XY}, \nu_{XY})$ be an optimal transport plan attaining $W(\mu_{XY}, \nu_{XY})$, with covariate marginal γ^*_{XX}. Assume that its conditional label distribution given $(X_1, X_2) = (x_1, x_2)$ satisfies $\gamma^*_{YY|x_1,x_2} \in \Gamma(\mu_{Y|x_1}, \nu_{Y|x_2})$ for γ^*_{XX}-almost every (x_1, x_2). Define the diagonal transport measure $\psi(A) := \gamma^*_{XX}\big(\{(x, x) : x \in A\}\big)$ for every measurable $A \subseteq \mathcal{X}$. Suppose there exists $c \in (0, 1]$ such that $\psi(A) \geq c\,\nu_X(A)$ for every measurable $A \subseteq \mathcal{X}$. Then defining $\eta := 1/c$, we can derive $\int_{\mathcal{X}} W(\mu_{Y|x}, \nu_{Y|x})\,\mathrm{d}\nu_X(x) \leq \eta \cdot W(\mu_{XY}, \nu_{XY})$.*

If P_{XY} and Q_{XY} satisfy the assumption of Theorem3.3, we replace $\mu_{XY} = P_{XY}$ and $\nu_{XY} = Q_{XY}$ to bound the integrated conditional Wasserstein distance by

$$\int_{\mathcal{X}} W(P_{Y|x}, Q_{Y|x})\mathrm{d}Q_X(x) \leq \eta \cdot W(P_{XY}, Q_{XY}). \tag{12}$$

Finally, based on Eq. (12), we deduce that

$$\mathrm{ICG}(P, Q) \leq \sqrt{2\kappa L}\,(\eta \cdot W(P_{XY}, Q_{XY}) + 1/4). \tag{13}$$

Eq. (13) relates distribution shift to conditional coverage robustness through an upper bound on ICG. For fixed κ, L, and η, reducing $W(P_{XY}, Q_{XY})$ tightens this bound, motivating alignment of the calibration and test distributions. Appendix A explains these constants and provides numerical bounds on η and empirical estimates of L. The finite-sample approximation of $W(P_{XY}, Q_{XY})$ is examined in Appendix B.

4 Method

The upper bound in Eq. (13) provides a framework to mitigate conditional-coverage degradation under distribution shift. Specifically, if a model f_θ transforms Q_{XY} via the Wasserstein transport plan to P_{XY}, for $(X_{n+1}, Y_{n+1}) \sim Q_{XY}$, we have

$$(\overline{X}_{n+1}, \overline{Y}_{n+1}) := f_\theta(X_{n+1}, Y_{n+1}) \sim P_{XY}. \tag{14}$$

Therefore, the conditional prediction set $\mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})$ constructed on the normalized input ensures conditional coverage with respect to P_{XY}. However, to achieve $1 - \alpha$ conditional coverage on Q_{XY} during inference, the model f_θ must satisfy two additional requirements.

(i) To obtain a prediction set of the original test input X_{n+1} from $\mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})$, f_θ should be invertible:

$$f_\theta^{-1}(\overline{X}_{n+1}, \overline{Y}_{n+1}) = (X_{n+1}, Y_{n+1}). \tag{15}$$

(ii) Since Y_{n+1} is unobserved at inference, f_θ should not depend on Y_{n+1} when transform of X_{n+1}:

$$\overline{X}_{n+1} \perp Y_{n+1} | X_{n+1}. \tag{16}$$

4.1 Branched Normalizing Flow

Normalizing flows are widely applied techniques for invertible mapping [25, 26]. A formal definition of normalizing flows is presented in Definition 4.1 with a demonstration in Figure 2.

Definition 4.1 (Normalizing flows [25])**.** Let μ_X be a probability measure in $\mathbb{R}^d$. For a measurable and invertible function $g : \mathbb{R}^d \to \mathbb{R}^d$, ν_X is the pushforward measure of μ_X through g, denoted as $\nu_X = g_{\#}\mu_X$, if $\nu_X(\mathcal{A}) = \mu_X(g^{-1}(\mathcal{A}))$ for every measurable set $\mathcal{A} \subseteq \mathbb{R}^d$. g is referred to as the generative flow, and $f = g^{-1}$ is known as the normalizing flow with $\mu_X = f_{\#}\nu_X$.

To make f_θ meet the two requirements (i) and (ii), we introduce a special normalizing flow, called **Branched Normalizing Flow (BNF)**. For a given sample (x, y), BNF transforms it with a branched structure such that

$$(\overline{x}, \overline{y}) = f_\theta(x, y) = (f_{\theta_X}(x), f_{\theta_Y}(y)). \quad (17)$$

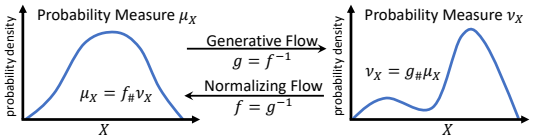


Figure 2: Invertible generative and normalizing flows.

The invertibility of BNF allows that $f_\theta^{-1}(\overline{x}, \overline{y}) = (f_{\theta_X}^{-1}(\overline{x}), f_{\theta_Y}^{-1}(\overline{y})) = (x, y)$, enabling the inverse transformation of $\mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})$ and satisfying requirement (i). Besides, the parameters θ_X and θ_Y are not shared between branches, so BNF does not explicitly couple the mappings of x and y. Therefore, the normalized test input $\overline{X}_{n+1}$ can be obtained without knowing Y_{n+1}, fulfilling requirement (ii).

Consider a BNF realizing $f_{\theta\#}Q_{XY} = P_{XY}$ so that $f_{\theta_X\#}Q_X = P_X$ and $f_{\theta_Y\#}Q_{Y|x} = P_{Y|\overline{x}}$ by optimizing

$$\min_\theta W(P_{XY}, f_{\theta\#}Q_{XY}). \quad (18)$$

Then, given a test input $X_{n+1} = x$, we normalize it as $\overline{X}_{n+1} := f_{\theta_X}(x) = \overline{x}$. Since $f_{\theta\#}Q_{XY} = P_{XY}$, the transformed true target $\overline{Y}_{n+1} := f_{\theta_Y}(Y_{n+1})$, together with $\overline{X}_{n+1}$, should follow the calibration distribution, i.e, $(\overline{X}_{n+1}, \overline{Y}_{n+1}) \sim P_{XY}$, as shown in Figure 1(c) 1st plot. Therefore, the adaptive prediction set of $\overline{X}_{n+1}$ ensures

$$\Pr\left(\overline{Y}_{n+1} \in \mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1}) | \overline{X}_{n+1} = \overline{x}\right) \geq 1 - \alpha. \quad (19)$$

BNF then constructs a prediction set of the original input X_{n+1} by including all targets whose normalized counterparts lie in $\mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})$. Specifically, we define

$$\mathcal{C}_{\mathrm{BNF}}(X_{n+1}) := \{f_{\theta_Y}^{-1}(\overline{y}) : \overline{y} \in \mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})\}. \quad (20)$$

Proposition C.1 in Appendix C and the invertibility of the univariate function f_{θ_Y} imply that

$$\overline{Y}_{n+1} \in \mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1}) \iff f_{\theta_Y}^{-1}(\overline{Y}_{n+1}) \in \mathcal{C}_{\mathrm{BNF}}(X_{n+1}). \quad (21)$$

Consequently, since $Y_{n+1} = f_{\theta_Y}^{-1}(\overline{Y}_{n+1})$, the conditional guarantee is inherited by $\mathcal{C}_{\mathrm{BNF}}(X_{n+1})$:

$$\Pr\left(Y_{n+1} \in \mathcal{C}_{\mathrm{BNF}}(X_{n+1}) | X_{n+1} = x\right) \geq 1 - \alpha. \quad (22)$$

Even if f_{θ_X} and f_{θ_Y} do not share parameters, Wasserstein minimization in Eq. (18) considers the dependency between them. We provide an illustrative example in Appendix D to show how this dependency is implicitly accounted for during optimization.

4.2 Enhancing Fitting Ability via Gaussian Noise Augmentation

The monotonicity of the univariate f_{θ_Y} allows the equivalence in Eq. (21), but also limits its fitting ability. As a result, it struggles to optimize Eq. (18) for complex distributions, leading to unreliable conditional coverage. We empirically present the issue in Appendix E.

To address this limitation, we adopt the augmentation technique proposed in [27] and introduce a variant called **Augmented BNF** to gain higher fitting ability. Specifically, given a sample (x, y), the augmented transformation is defined as

$$(\overline{x}, \overline{y}) := (f_{\theta_X}(x), f_{\theta_Y}^{\mathrm{aug}}(y; \varepsilon)). \quad (23)$$

where ε is sampled from a Gaussian distribution $\mathcal{N}(0, 1)$. Meanwhile, f_{θ_X} is unchanged from BNF. We implement Augmented BNF using Real NVP [28, 27], a representative coupling flow with architectural details and computational costs provided in Appendix F.

Although the full transformation $f_{\theta_Y}^{\mathrm{aug}}(y; \varepsilon)$ remains invertible, it does not build a monotonic relationship between y and $\overline{y}$. As a result, we cannot rely on Proposition C.1 to preserve the conditional guarantee. To address this issue, we propose an alternative approach to obtain a prediction set for test input X_{n+1} with a sampled noise ε_{n+1} by defining

$$\mathcal{C}_{\mathrm{BNF}}^{\mathrm{aug}}(X_{n+1}) := \left\{y : f_{\theta_Y}^{\mathrm{aug}}(y; \varepsilon_{n+1}) \in \mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})\right\}. \quad (24)$$

Proposition C.2 in Appendix C implies that $\overline{Y}_{n+1} \in \mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1}) \iff Y_{n+1} \in \mathcal{C}^{\mathrm{aug}}_{\mathrm{BNF}}(X_{n+1})$. Hence, based on Eq. (19), after averaging over the augmentation noise ε_{n+1}, we conclude that

$$\Pr\left(Y_{n+1} \in \mathcal{C}^{\mathrm{aug}}_{\mathrm{BNF}}(X_{n+1}) | X_{n+1} = x\right) \geq 1 - \alpha. \tag{25}$$

Eq. (19), (22), and (25) assume that the transformed samples follow P_{XY} exactly and that $\mathcal{C}_C$ satisfies the conditional coverage guarantee in Eq. (3). With finite samples, estimating $\mathcal{C}_C$ introduces approximation error, and the learned transport may leave residual distribution mismatch. Appendix M.2 analyzes these two errors in detail.

5 Application to Multi-Source Domains

In this work, we study joint distribution shift in multi-source domain generalization (MSDG) [29], a widely explored setting in CP [19, 20, 8]. In MSDG, the test distribution is a random mixture within the convex hull of the source distributions. Formally, given K source distributions D^k_{XY} for k=1, .., K, the test distribution satisfies $Q_{XY} \in \left\{\sum_{k=1}^K \lambda_k D^k_{XY} : \lambda_1, ..., \lambda_K \geq 0, \sum_{k=1}^K \lambda_k = 1\right\}$.

Theorem 5.1. *Let $\{\nu^k_{XY}\}_{k=1}^K$ be probability measures defined on the metric space $(\mathcal{X} \times \mathcal{Y}, d_{\mathcal{XY}})$, and let ν_{XY} lie in the convex hull of these measures, i.e., $\nu_{XY} = \sum_{k=1}^K \lambda_k \nu^k_{XY}$ with $\lambda_k \geq 0$ and $\sum_{k=1}^K \lambda_k = 1$. For any probability measure μ_{XY} on $(\mathcal{X} \times \mathcal{Y}, d_{\mathcal{XY}})$, the following inequality holds: $W(\mu_{XY}, \nu_{XY}) \leq \sum_{k=1}^K \lambda_k W(\mu_{XY}, \nu^k_{XY})$.*

As outlined in [19, 8], achieving coverage guarantee for each source distribution ensures that the coverage on the test distribution is preserved. Inspired by the principle, Theorem 5.1 suggests a surrogate objective for Augmented BNF by $\sum_{k=1}^K \lambda_k W(P_{XY}, {f^{\mathrm{aug}}_\theta}_{\#} D^k_{XY})$. Since $\{\lambda_k\}_{k=1}^K$ are unknown, we minimize the expectation assuming they are uniformly distributed over the simplex:

$$\min_\theta \tfrac{1}{K} \sum\nolimits_{k=1}^K W(P_{XY}, {f^{\mathrm{aug}}_\theta}_{\#} D^k_{XY}). \tag{26}$$

We work with finite samples in practice. Let S_{D^k} denote equally sized training sets drawn from the source distributions D^k_{XY}, for $k = 1, \ldots, K$. We take the calibration distribution to be the equally weighted mixture of the source distributions, $P_{XY} := \frac{1}{K}\sum_{k=1}^K D^k_{XY}$. Given equally sized source training sets S_{D^k}, their union $S^{\mathrm{align}}_P := \bigcup_{k=1}^K S_{D^k}$ provides an empirical approximation of P_{XY}, which we use for training Augmented BNF. We draw a separate calibration set S^{cal}_P from the same distribution P_{XY}, disjoint from S^{align}_P, and reserve it exclusively for conformal calibration. During optimization, for each $(x, y) \in \mathcal{S}_{D^k}$, we sample a noise $\varepsilon \sim \mathcal{N}(0, 1)$ and compute $(\bar{x}, \bar{y})$ using Eq. (23). All normalized pairs are collected in $\bar{\mathcal{S}}_{D^k}$. We estimate empirical distributions $\widehat{P}_{XY}$ and ${f^{\mathrm{aug}}_\theta}_{\#} \widehat{D}^k_{XY}$ using $\mathcal{S}^{\mathrm{align}}_P$ and $\bar{\mathcal{S}}_{D^k}$, respectively, to optimize the objective in Eq. (26).

Even if $(\overline{X}_{n+1}, \overline{Y}_{n+1}) \sim P_{XY}$, constructing $\mathcal{C}_{\mathrm{C}}(\overline{X}_{n+1})$ that satisfies the conditional guarantee under P_{XY} remains challenging with finite samples [18]. In this work, we employ conformalized quantile regression (CQR) [30], which generates a prediction set $\mathcal{C}_{\mathrm{CQR}}(\overline{X}_{n+1})$ to approximate the $1 - \alpha$ conditional coverage in Eq. (19). Crucially, CQR is independent of the

Algorithm 1 Augmented BNF + CQR under MSDG

Require: sets $\mathcal{S}_{D^k}$ for k=1, ..., K; alignment set $\mathcal{S}^{\mathrm{align}}_P$, calibration set $\mathcal{S}^{\mathrm{cal}}_P$; test set $\mathcal{S}_Q$; N epochs; $1-\alpha$ confidence; Augmented BNF f^{aug}_θ; CQR algorithm A_{CQR}.

Training Phase:
for i=1 to N epochs **do**
  for k=1 to K **do**
    Initialize $\bar{\mathcal{S}}_{D^k} \leftarrow \emptyset$
    for each $(x, y) \in \mathcal{S}_{D^k}$ **do**
      $(\bar{x}, \bar{y})$=$f^{\mathrm{aug}}_\theta(x, y, \varepsilon)$, where $\varepsilon \sim \mathcal{N}(0, 1)$
      $\bar{\mathcal{S}}_{D^k} \leftarrow \bar{\mathcal{S}}_{D^k} \cup \{(\bar{x}, \bar{y})\}$
    end for
  end for
  $\min_\theta \frac{1}{K}\sum_{k=1}^K W\left(\widehat{P}_{XY}, {f^{\mathrm{aug}}_\theta}_{\#} \widehat{D}^k_{XY}\right)$
end for

Inference Phase:
for x from $\mathcal{S}_Q$ **do**
  $\bar{x}$=$f_{\theta_X}(x)$
  $\mathcal{C}_{\mathrm{CQR}}(\bar{x})$=$A_{\mathrm{CQR}}\left(\bigcup_{k=1}^K \mathcal{S}_{D^k}, \mathcal{S}^{\mathrm{cal}}_P, \bar{x}, 1-\alpha\right)$
  Sample $\varepsilon \sim \mathcal{N}(0, 1)$
  $\mathcal{C}^{\mathrm{aug}}_{\mathrm{BNF}}(x)$=$\{y : f^{\mathrm{aug}}_{\theta_Y}(y; \varepsilon) \in \mathcal{C}_{\mathrm{CQR}}(\bar{x})\}$
end for

Augmented BNF and can be seamlessly integrated. Appendix G provides the details of CQR, whose implementation is indicated as $A_{\text{CQR}}(\bigcup_{k=1}^{K} \mathcal{S}_{D^k}, \mathcal{S}_P^{\text{cal}}, \overline{X}_{n+1}, 1-\alpha)$. The pseudo-code of Augmented BNF + CQR is outlined in Algorithm 1.

6 Experiment

6.1 Experimental Setup

We conduct Augmented BNF using the normflows library [31]. To estimate the empirical Wasserstein distance, we adopt the Sinkhorn algorithm [32, 33] via the geomloss library [34] in Appendix H. The implementation of our method is publicly available.[3]

Baselines. Five methods are selected for comparison. Split CP (SCP) [15] ensures marginal coverage under i.i.d. data; Importance-Weighted CP (IW-CP) [35] addresses covariate shift; Worst-Case CP (WC-CP) [19, 20, 36] provides conservative guarantees under joint distribution shift; and Wasserstein-Regularized CP (WR-CP) [8] improves robustness under MSDG. We also include CQR alone, without Augmented BNF, to highlight its limitations in pursuing conditional coverage under shift. Additional details about the baselines are given in Appendix I, with examples in Figure 12.

Datasets. We set $K = 3$ under both **synthetic** and **natural** distribution shifts. Synthetic shifts are introduced in the PTS dataset [9]. For real-world applications, we consider (i) sales prediction over time with Bike Rental [10], (ii) multi-location traffic forecasting with Seattle-Loop [11], PEMSD4, and PEMSD8 [37], (iii) unbiased healthcare with MIMIC-IV [12], eICU [13], and data from a collaborating hospital, and (iv) epidemic modeling across pandemic phases with U.S. Influenza-like Illness (ILI) [14]. Dataset summaries and data-preparation details are provided in Appendix J.

Evaluation metric. The worst-slice coverage (WSC) [38] measures the minimal coverage over any sufficiently large slice in $\mathcal{X}$, serving as an empirical proxy for the robustness of conditional coverage. However, as reviewed in Appendix K, WSC captures only the minimal (i.e., most insufficient) coverage and overlooks regions where coverage may be overly conservative. To address these weaknesses, we propose **worst-slice coverage gap (WSCG)** that captures both under- and over-coverage by taking the maximum absolute deviation from $1-\alpha$ over slices containing at least 10% of test samples. Specifically, for any CP methods that produce a prediction set $\mathcal{C}(x)$ given an input x,

$$\text{WSCG} = \sup_{\mathcal{S}\subseteq\mathcal{X}} \left|\Pr(y \in \mathcal{C}(x) | x \in \mathcal{S}) - (1-\alpha)\right|, \quad \text{s.t.}\, \Pr(x \in \mathcal{S} | (x,y) \in \mathcal{S}_Q) \geq 0.1. \quad (27)$$

6.2 Main Result

We evaluate the combination of Augmented BNF and CQR, along with five baseline methods, across 10 independent trials for each dataset. For each trial, 100 random mixtures were generated as test sets. Coverage metrics are evaluated on all 100 test sets per trial. The results are summarized in Figure 3 with $1-\alpha = 0.9$. Our approach consistently achieves marginal coverage and WSC values close to the desired confidence level, and obtains the lowest WSCG. Additional experiments examine performance across sample sizes, K, and confidence levels (Appendices L and O.1) and show that Augmented BNF reduces WSCG with both SCP and CQR (Appendix O.2).

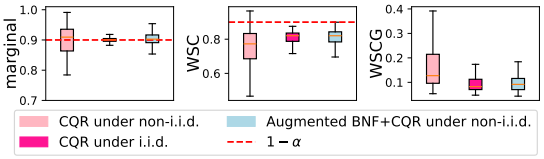


Figure 4: Comparison of Augmented BNF + CQR under distribution shift with CQR in shifted and i.i.d. settings.

As discussed at the end of Section 4.2, our finite-sample implementation involves both conditional-coverage approximation error and imperfect transformation. Consequently, Augmented BNF + CQR may retain nonzero WSCG and deviations from the target marginal coverage, as observed in Figure 3. To quantify the resulting coverage errors empirically, Figure 4 compares Augmented BNF + CQR under distribution shift with CQR in both shifted and i.i.d. settings. The i.i.d. CQR results provide a reference for the conditional-coverage approximation error, with nonzero WSCG even without distribution shift. Our method closely approaches this reference, indicating that the learned transformation mitigates the additional coverage error associated with distribution shift. Appendix M.1 further derives a marginal coverage lower bound under imperfect transformations.

[3] https://github.com/rxu0112/Branched-normalizing-flow

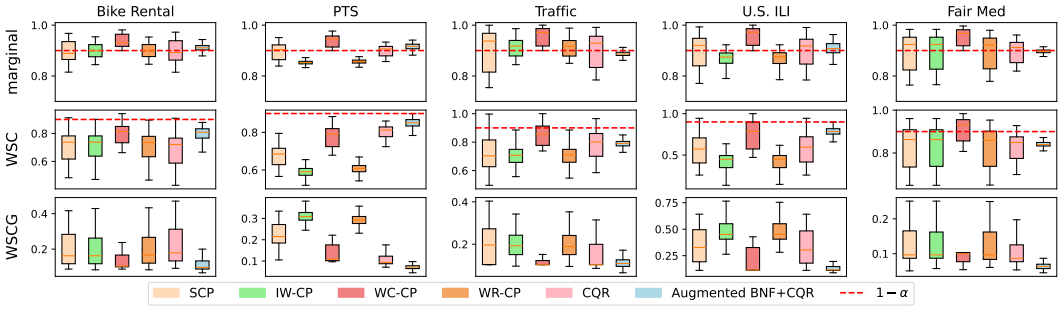


Figure 3: Marginal coverage, WSC, and WSCG of Augmented BNF+CQR and baselines with $1 - \alpha = 0.9$: our method achieves the lowest WSCG and brings the marginal coverage and WSC close to the expected confidence.

7 Discussion

7.1 Feature vs. Stochastic Conditioning in BNF

A natural extension of the original BNF in Eq. (17) is to condition the Y-transformation on features, denoted by $f^{\text{fea}}_{\theta_Y}(y; x)$, in order to explicitly capture input dependence. We refer to this variant as **Feature-Conditioned BNF**. Under exact joint alignment of Q_{XY} with P_{XY}, the following prediction set inherits the population conditional coverage of $\mathcal{C}_C$ in Eq. (3), as in Eq. (25):

$$\mathcal{C}^{\text{fea}}_{\text{BNF}}(x) := \left\{ y : f^{\text{fea}}_{\theta_Y}(y; x) \in \mathcal{C}_C(f_{\theta_X}(x)) \right\},$$

$$\Pr\left(Y_{n+1} \in \mathcal{C}^{\text{fea}}_{\text{BNF}}(X_{n+1}) \mid X_{n+1} = x\right) \geq 1 - \alpha.$$

However, feature conditioning reduces the sample-to-dimension ratio, defined as the number of training observations divided by the total input dimension of the two branches, from $|S_{D^k}|/(d+2)$ for Augmented BNF to $|S_{D^k}|/(2d+1)$ for Feature-Conditioned BNF, where d is the feature dimension, as detailed in Table 4. This can make distribution estimation more difficult, consistent with the less robust conditional coverage observed in Figure 5.

In contrast, $f^{\text{aug}}_{\theta_Y}(y; \varepsilon)$ in Augmented BNF can be viewed as conditioning on a simple one-dimensional Gaussian noise variable ε. Although this design leaves the learning of feature dependence entirely to the joint Wasserstein minimization in Eq. (18), the injected noise ε effectively introduces latent degrees of freedom, turning $f^{\text{aug}}_{\theta_Y}$ into a stochastic mapping. This allows the model to represent *a distribution over transformations*, thereby substantially increasing expressiveness beyond a deterministic map.

Replacing the original input x of $f^{\text{fea}}_{\theta_Y}(y; x)$ with a one-dimensional representation $\tilde{x}$ (e.g., via PCA [39], t-SNE [40], or UMAP [41]) can alleviate the curse of dimensionality. Nevertheless, as a deterministic and compressed projection, $\tilde{x}$ inevitably loses information from the original input and lacks the flexibility of stochastic conditioning. Consequently, it is less expressive than ε and less informative than x, leading to inferior performance in Figure 5. The results for each dataset are shown in Appendix O.3.

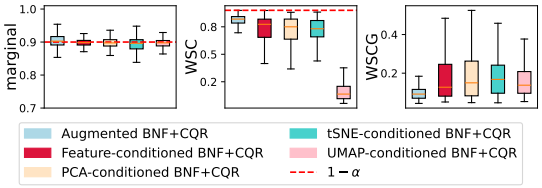


Figure 5: Comparison of conditioning the transformation of Y on the feature and its one-dimensional projection. The result is aggregated over datasets.

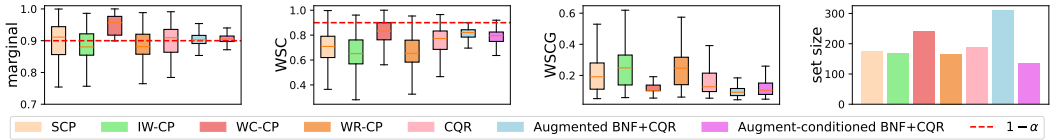


Figure 6: Standard Augmented BNF can produce large prediction sets under MSDG. The Augment-Conditioned variant significantly reduces set size and preserves coverage performance. The result is aggregated over datasets.

7.2 Enhancing Prediction Efficiency with Source-Conditioned Transformations

Prediction efficiency in CP is typically measured by prediction set size, with smaller sets being more informative at a fixed coverage level. In Augmented BNF, however, the sampled noise ε_{n+1} in Eq. (24) is source-agnostic and provides no explicit source information to $f^{\text{aug}}_{\theta_Y}$. Accommodating different source conditional label distributions through a shared transformation can therefore lead to larger prediction sets.

To improve efficiency, we introduce **Augment-Conditioned BNF**, which incorporates source-specific conditioning into the transformation. As shown in Figure 6, this yields smaller prediction sets while preserving coverage robustness. For both BNF variants, prediction set size is estimated using 10 test sets per trial, totaling 100 test sets across 10 independent trials for each dataset (see Appendix F.2). Appendix N provides methodological details and results for individual datasets.

7.3 Distribution Shift as Label Perturbation

We also evaluate our method under multiplicative label perturbations [42, 43]. The perturbed label is $\widetilde{Y} = \delta Y$, with $\delta \in [1, 1.5]$, while X remains unchanged. Thus, $Q_X = P_X$, but the conditional label distribution changes ($Q_{Y|X} \neq P_{Y|X}$). This setting examines whether BNF can correct label-distribution mismatch when the feature marginals provide no indication of the shift. We construct multiple such shifted environments and train BNF to transport them back to the unperturbed calibration distribution. During inference, we evaluate on 100 randomly shifted distributions.

Figure 7 shows that Augmented BNF+CQR maintains marginal coverage close to the target and competitive WSCG, while producing substantially smaller prediction sets than WC-CP in the pooled comparison. The label perturbation setup and results for each dataset are detailed in Appendix O.5.

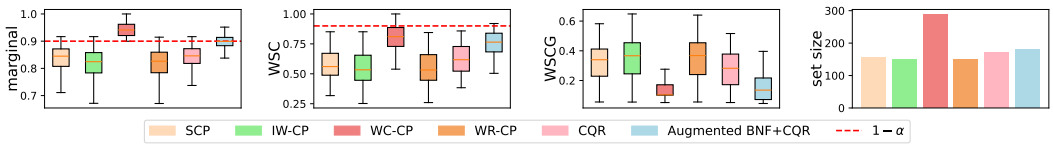


Figure 7: Distribution shift is induced by label perturbation. Our method consistently maintains comparably robust coverage without a significant cost in prediction efficiency. The result is aggregated over datasets.

7.4 Scope and Limitations

Augmented BNF improves fitting flexibility through stochastic augmentation, but this also makes prediction sets in Eq. (24) depend on the sampled noise. Repeated evaluations of the same input may therefore yield different sets, which can complicate applications requiring consistent decisions. Fixing the random seed can make repeated predictions reproducible.

Beyond this stochasticity, the current RealNVP feature branch uses coupling layers that require at least two input dimensions, restricting its direct application to tasks with multivariate features. For univariate inputs, one possible adaptation is to augment the feature representation with noise or engineered features. Appendix P discusses these limitations and possible remedies.

8 Conclusion

This work proposes the Conditional Coverage Gap (CCG) to evaluate the robustness of conditional coverage at a given test input, and defines the Integrated Coverage Gap (ICG) as its expectation over the test feature distribution. We bound ICG using the Wasserstein distance $W(P_{XY}, Q_{XY})$, capturing how distribution shift propagates from the data space to the conformal score space. To improve conditional-coverage robustness under shift, we introduce the Branched Normalizing Flow (BNF). The invertibility of BNF enables mapping adaptive prediction sets from P_{XY} to Q_{XY}, while the branched structure allows input X_{n+1} transformation without needing Y_{n+1} at test time. BNF is applied to both synthetic and real-world distribution shifts to validate its effectiveness.

9 Acknowledgement

This work was supported in part by the National Key R&D Program of China (Grant No. 2023YFF0725001), the National Natural Science Foundation of China (Grant No. 92370204), the Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence, the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515120057), the Key-Area Special Project of Guangdong Provincial Ordinary Universities (Grant No. 2024ZDZX1007), the Department of Science and Technology of Guangdong Province (Grant No. 2023CX10X079), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), and the Education Bureau of Guangzhou Municipality.

References

[1] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[2] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction, 2007.

[3] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017.

[4] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

[5] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

[6] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[7] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

[8] Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Prashant Rana. Physicochemical Properties of Protein Tertiary Structure. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5QW3H.

[10] Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5W894.

[11] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[12] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[13] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[14] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 245–254, 2020.

[15] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

[16] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

[17] 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.

[18] Rina Foygel Barber, Emmanuel J Candes, 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.

[19] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pages 1–66, 2024.

[20] Xin Zou and Weiwei Liu. Coverage-guaranteed prediction sets for out-of-distribution data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17263–17270, 2024.

[21] Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR, 2023.

[22] Haifeng Wen, Hong Xing, and Osvaldo Simeone. Distributed conformal prediction via message passing, 2025.

[23] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.

[24] Nathan Ross. Fundamentals of stein's method. 2011.

[25] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[26] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[27] Chin-Wei Huang, Laurent Dinh, and Aaron Courville. Augmented normalizing flows: Bridging the gap between generative flows and latent variable models, 2020.

[28] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[29] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[30] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[31] Vincent Stimper, David Liu, Andrew Campbell, Vincent Berenz, Lukas Ryll, Bernhard Schölkopf, and José Miguel Hernández-Lobato. normflows: A pytorch package for normalizing flows. *Journal of Open Source Software*, 8(86):5361, 2023.

[32] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[33] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[34] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

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

[36] Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

[37] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

[38] 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.

[39] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[41] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[42] Matteo Sesia, YX Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *arXiv preprint arXiv:2309.05092*, 2023.

[43] Bat-Sheva Einbinder, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to label noise. *arXiv preprint arXiv:2209.14295*, 2, 2022.

[44] Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[45] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.

[46] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[48] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

[49] Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 24–38. PMLR, 2021.

[50] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Locally valid and discriminative prediction intervals for deep learning models. *Advances in Neural Information Processing Systems*, 34:8378–8391, 2021.

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

[52] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

[53] Alvaro Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. *Advances in Neural Information Processing Systems*, 37:101000–101041, 2024.

[54] David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

[55] Batiste Le Bars and Pierre Humbert. On volume minimization in conformal regression, 2025.

[56] Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34:2060–2071, 2021.

[57] Nicolo Colombo. Normalizing flows for conformal regression. *arXiv preprint arXiv:2406.03346*, 2024.

[58] Zhenhan Fang, Aixin Tan, and Jian Huang. Contra: Conformal prediction region via normalizing flow transformation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[59] Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport, 2025.

[60] Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction, 2025.

[61] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. 2011.

[62] Danijel Kivaranovic, Kory D. Johnson, and Hannes Leeb. Adaptive, distribution-free prediction intervals for deep networks, 2020.

[63] Matteo Sesia and Emmanuel J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1), January 2020.

[64] Jean Feydy. *Analyse de données géométriques, au delà des convolutions*. PhD thesis, Université Paris-Saclay, 2020.

[65] Xiaoyong Jin, Youngsuk Park, Danielle Maddix, Hao Wang, and Yuyang Wang. Domain adaptation for time series forecasting via attention sharing. In *International Conference on Machine Learning*, pages 10280–10297. PMLR, 2022.

[66] Yue Sun, Chao Chen, Yuesheng Xu, Sihong Xie, Rick S Blum, and Parv Venkitasubramaniam. Reaction-diffusion graph ordinary differential equation networks: Traffic-law-informed speed prediction under mismatched data. The 12th International Workshop on Urban Computing, held in conjunction with . . . , 2023.

[67] Manoja Kumar Das. Multicenter studies: relevance, design and implementation. *Indian pediatrics*, 59(7):571–579, 2022.

[68] Henrik Olsson, Kimmo Kartasalo, Nita Mulliqi, Marco Capuccini, Pekka Ruusuvuori, Hemamali Samaratunga, Brett Delahunt, Cecilia Lindskog, Emiel AM Janssen, Anders Blilie, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature communications*, 13(1):7761, 2022.

[69] CDC. Pandemic Intervals Framework (PIF) — cdc.gov. `https://www.cdc.gov/pandemic-flu/php/national-strategy/intervals-framework.html?CDC_AAref_Val=https://www.cdc.gov/flu/pandemic-resources/national-strategy/intervals-framework.html`. [Accessed 28-04-2025].

[70] Tiberiu Harko, Francisco SN Lobo, and Man Kwong Mak. Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation*, 236:184–194, 2014.

[71] KM Ariful Kabir, Kazuki Kuga, and Jun Tanimoto. Analysis of sir epidemic model with information spreading of awareness. *Chaos, Solitons & Fractals*, 119:118–125, 2019.

[72] Mustafa Turkyilmazoglu. A restricted epidemic sir model with elementary solutions. *Physica A: Statistical Mechanics and its Applications*, 600:127570, 2022.

[73] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[74] Zhifeng Kong and Kamalika Chaudhuri. The expressive power of a class of normalizing flow models, 2020.

[75] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

A Interpretation and Quantification of the Scaling Constants

A.1 Roles of κ, L, and η in the ICG Bound

We explain the roles of the scaling constants in Eq. (13) and provide their quantitative characterization for the score function and distribution shifts considered in our experiments.

First, the term L, representing the Lebesgue density bound of $P_{V|x}$, captures the concentration of conformal scores at x. A higher L indicates that the calibration scores are tightly clustered around certain values of v, which makes the conditional coverage more sensitive to distribution shifts in test conformal scores. Hence, this highlights how the shape of the calibration conformal score distribution directly influences coverage robustness.

Second, κ provides an interpretation of how the score function $s(x, y)$ influences robustness under distribution shift. Specifically, the continuity constant $\kappa \geq (|s(x, y_1) - s(x, y_2)|)/(|y_1 - y_2|)$ for all $y_1, y_2 \in \mathcal{Y}, x \in \mathcal{X}$. It captures the sensitivity of the score function $s(x, y)$ to changes in the label y, given a fixed input x. A smaller κ implies that the conformal score is relatively insensitive to variations in the label, meaning that even under a large concept shift (i.e., large $W(P_{Y|x}, Q_{Y|x})$), the induced shift in conformal scores $W(P_{V|x}, Q_{V|x})$ remains small.

Lastly, $\eta = 1/c$ is determined by the diagonal-overlap constant in Theorem 3.3. A larger c gives a smaller η and therefore tightens the bound in Eq. (13).

A.2 Characterization of κ and L for CQR

For the CQR score introduced in Appendix G, $s(x, y) = \max\{h^{\mathrm{lo}}(x) - y,\ y - h^{\mathrm{hi}}(x)\}$, both terms are 1-Lipschitz in y. Therefore, $|s(x, y_1) - s(x, y_2)| \leq |y_1 - y_2|$, and Theorem 3.2 applies with $\kappa = 1$.

Assume that $P_{Y|x}$ admits a density, denoted by $p_{Y|x}$, and let $p_{V|x}$ denote the density of the corresponding CQR score under P_{XY}. For $v > (h^{\mathrm{lo}}(x) - h^{\mathrm{hi}}(x))/2$,

$$p_{V|x}(v) = p_{Y|x}\big(h^{\mathrm{lo}}(x) - v\big) + p_{Y|x}\big(h^{\mathrm{hi}}(x) + v\big), \tag{28}$$

while the density is zero below this threshold. If the conditional label densities are uniformly bounded, we may choose

$$L := \sup_{x,v} p_{V|x}(v) \leq 2 \sup_{x,y} p_{Y|x}(y). \tag{29}$$

We numerically estimate L from the calibration CQR scores. We first standardize the features and form a local neighborhood around each calibration input. Let m_{nn} denote the neighborhood size, $\sigma_{i,\mathrm{loc}}$ the sample standard deviation of the scores in the neighborhood of x_i, and σ_{all} the standard deviation of all calibration scores. The adaptive Gaussian-KDE bandwidth used for x_i is

$$\mathrm{bw}_i = \max\left\{1.06\, \sigma_{i,\mathrm{loc}}\, m_{\mathrm{nn}}^{-1/5}, \max\{10^{-3}\sigma_{\mathrm{all}}, 10^{-6}\}\right\}.$$

For each neighborhood, we evaluate the density on a 256-point grid spanning its observed score range and define $\widehat{L}$ as the largest density value over all neighborhoods and grid points. To examine the sensitivity of $\widehat{L}$ to the neighborhood size, we report results for $m_{\mathrm{nn}} \in \{50, 100, 150, 200\}$ in Table 1.

Table 1: Empirical estimates of $\widehat{L}$ at $1 - \alpha = 0.9$ for different neighborhood sizes.

Setting	$m_{\mathrm{nn}} = 50$	$m_{\mathrm{nn}} = 100$	$m_{\mathrm{nn}} = 150$	$m_{\mathrm{nn}} = 200$
Bike Rental	0.0150	0.0145	0.0120	0.0111
PTS	0.9932	0.6231	0.5470	0.4792
Traffic	1.4625	0.9639	0.8882	0.8305
U.S. ILI	0.1021	0.0664	0.0382	0.0313
Fair Med	0.1390	0.1356	0.1270	0.1166

A.3 Bounds on η under the considered shifts

Multi-source domain generalization. Write the calibration and test distributions as

$$P_{XY} = \sum_{k=1}^{K} \pi_k D_{XY}^k, \qquad Q_{XY} = \sum_{k=1}^{K} \lambda_k D_{XY}^k,$$

where $\pi_k > 0$, $\lambda_k \geq 0$, and $\sum_k \pi_k = \sum_k \lambda_k = 1$. Here, π_k and λ_k are the calibration and test mixture weights, respectively, and D_X^k denotes the covariate marginal of D_{XY}^k.

Suppose that an optimal plan satisfying the conditional-coupling assumption of Theorem 3.3 has a diagonal transport measure satisfying

$$\psi(A) \geq \sum_{k=1}^{K} \min(\pi_k, \lambda_k) D_X^k(A)$$

for every measurable $A \subseteq \mathcal{X}$. This condition bounds the mass transported without changing the covariate using the shared mixture weights. Since $\max_k(\lambda_k/\pi_k) \geq 1$, we have

$$\min(\pi_k, \lambda_k) \geq \frac{\lambda_k}{\max_j(\lambda_j/\pi_j)}.$$

Consequently,

$$\psi(A) \geq \frac{1}{\max_k(\lambda_k/\pi_k)} Q_X(A).$$

The constant c in Theorem 3.3 can therefore be chosen to satisfy

$$c \geq \frac{1}{\max_k(\lambda_k/\pi_k)}, \qquad \eta = \frac{1}{c} \leq \max_k \frac{\lambda_k}{\pi_k}.$$

For an equally weighted calibration mixture with $K = 3$ and $\pi_k = \frac{1}{3}$, this gives $\eta \leq 3 \max_k \lambda_k \leq 3$. When the test weights are sampled uniformly over the three-source simplex and the stated transport condition holds for these mixtures,

$$\mathbb{E}[\eta] \leq 3\,\mathbb{E}[\max_k \lambda_k] = \frac{11}{6}.$$

Label perturbation. In Section 7.3, an environment with scaling factor $\delta \in [1, 1.5]$ has shifted distribution given by $(X, \delta Y)$, where $(X, Y) \sim P_{XY}$. Consider the coupling that pairs (X, Y) with $(X, \delta Y)$. Under the 2-product metric with $d_{\mathcal{Y}}(y_1, y_2) = |y_1 - y_2|$, its cost is

$$\mathbb{E}[|\delta Y - Y|] = \mathbb{E}[(\delta - 1)|Y|].$$

Every coupling between P_{XY} and Q_{XY} has cost at least

$$\mathbb{E}[|\delta Y|] - \mathbb{E}[|Y|] = \mathbb{E}[(\delta - 1)|Y|],$$

because the joint transport distance is at least the difference between the absolute label values. Thus, the proposed coupling is optimal. It also satisfies the conditional-coupling assumption of Theorem 3.3 and keeps every covariate unchanged. Hence, $\psi = P_X = Q_X$, $c = 1$, and $\eta = 1$.

B Finite-sample approximation of Wasserstein distance

In practice, the population forms of calibration and test distributions are typically inaccessible, so we may approximate $W(P_{XY}, Q_{XY})$ based on empirical distributions.

Definition B.1 (p-Wasserstein Distance between Empirical Distributions [23])**.** Let $\{x_i\}_{i=1}^n \sim \mu_X$ and $\{x'_j\}_{j=1}^m \sim \nu_X$ be i.i.d. samples from two distributions on a metric space $(\mathcal{X}, d_{\mathcal{X}})$. The Dirac measure ε_x is the point mass at $x \in \mathcal{X}$. The empirical measures are defined as

$$\widehat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{x_i}, \quad \widehat{\nu}_X = \frac{1}{m} \sum_{j=1}^{m} \varepsilon_{x'_j}.$$

$C \in \mathbb{R}^{n\times m}$ is a cost matrix where each element $C_{ij} = d_{\mathcal{X}}(x_i, x'_j)$ measures the distance between sample x_i from $\widehat{\mu}_X$ and x'_j from $\widehat{\nu}_X$. Let $\gamma \in \mathbb{R}^{n\times m}$ be a transportation plan matrix, where each $\gamma_{ij} \geq 0$ represents the mass transported from x_i to x'_j. The set of admissible transport plans is

$$\Gamma(\widehat{\mu}_X, \widehat{\nu}_X) = \left\{ \gamma \in \mathbb{R}^{n\times m}_{\geq 0} \;\middle|\; \sum_{j=1}^{m} \gamma_{ij} = \frac{1}{n},\ \sum_{i=1}^{n} \gamma_{ij} = \frac{1}{m} \right\}.$$

The p-Wasserstein distance between empirical distributions $\widehat{\mu}_X$ and $\widehat{\nu}_X$ is then given by

$$W_p(\widehat{\mu}_X, \widehat{\nu}_X) = \left(\min_{\gamma\in\Gamma(\widehat{\mu}_X,\widehat{\nu}_X)} \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij}\, C_{ij}^p \right)^{1/p}.$$

Let $\widehat{P}_{XY}$ and $\widehat{Q}_{XY}$ be the empirical distributions based on n and m i.i.d. samples drawn from P_{XY} and Q_{XY}, respectively. Our goal is to bound the deviation between the empirical and population Wasserstein distances, i.e., to analyze how $W(\widehat{P}_{XY}, \widehat{Q}_{XY})$ converges to $W(P_{XY}, Q_{XY})$ as n increases.

Definition B.2 (Upper Wasserstein Dimension [44])**.** Given a set $\mathcal{A} \subseteq \mathcal{X}$, the ϵ-covering number, denoted $\mathcal{N}_\epsilon(\mathcal{A})$, is the smallest n such that n closed balls, $\mathcal{U}_1, ..., \mathcal{U}_n$, of diameter ϵ achieve $\mathcal{A} \subseteq \cup_{1\leq i\leq n}\mathcal{U}_i$. For a distribution μ_X in $\mathcal{X}$, the (ϵ, ζ)-dimension is $d_\epsilon(\mu_X, \zeta) = -\log(\inf\{\mathcal{N}_\epsilon(\mathcal{A}) : \mu_X(\mathcal{A}) \geq 1-\zeta\})/\log\epsilon$. The upper Wasserstein dimension with $p = 1$ is

$$d_W(\mu_X) = \inf\{\varphi \in (2,\infty) : \limsup_{\epsilon\to 0} d_\epsilon(\mu_X, \epsilon^{\frac{\varphi}{\varphi-2}}) \leq \varphi\}. \tag{30}$$

Theorem B.3. *Given a probability measure μ_X in space $\mathcal{X}$, let $\sigma > d_W(\mu_X)$. If $\widehat{\mu}_X$ is an empirical measure corresponding to n i.i.d. samples from μ_X, $\exists\lambda \in \mathbb{R}$ such that $\mathbb{E}[W(\mu_X, \widehat{\mu}_X)] \leq \lambda n^{-1/\sigma}$. Furthermore, for $t > 0$,* $\Pr(W(\mu_X, \widehat{\mu}_X) \geq \mathbb{E}[W(\mu_X, \widehat{\mu}_X)] + t) \leq e^{-2nt^2}$ *[45].*

Theorem B.4. *Given probability measures μ_X and ν_X in space $\mathcal{X}$, let $\sigma_\mu > d_W(\mu_X)$ and $\sigma_\nu > d_W(\nu_X)$. Let $\widehat{\mu}_X$ and $\widehat{\nu}_X$ be the empirical measures constructed from two independent samples, consisting of n i.i.d. observations from μ_X and m i.i.d. observations from ν_X, respectively. For $t_\mu, t_\nu > 0$, $\exists\lambda_\mu, \lambda_\nu > 0$ with probability at least $(1 - e^{-2nt_\mu{}^2})(1 - e^{-2mt_\nu{}^2})$ that*

$$|W(\mu_X, \nu_X) - W(\widehat{\mu}_X, \widehat{\nu}_X)| \leq \lambda_\mu n^{-1/\sigma_\mu} + \lambda_\nu m^{-1/\sigma_\nu} + t_\mu + t_\nu. \tag{31}$$

A related theorem is proposed in [8], though without accounting for the signs of λ_μ and λ_ν. Based on Theorem B.4, if $\sigma_P > d_W(P_{XY})$ and $\sigma_Q > d_W(Q_{XY})$, for $t_P, t_Q > 0$, there are $\lambda_P, \lambda_Q > 0$ with a probability at least $(1 - e^{-2nt_P^2})(1 - e^{-2mt_Q^2})$ that

$$\left|W(P_{XY}, Q_{XY}) - W(\widehat{P}_{XY}, \widehat{Q}_{XY})\right| \leq \lambda_P n^{-1/\sigma_P} + \lambda_Q m^{-1/\sigma_Q} + t_P + t_Q. \tag{32}$$

As n and m increase, the bound in Eq. (32) decreases, thereby improving the approximation of the empirical Wasserstein distance. At the same time, the probability $(1 - e^{-2nt_P^2})(1 - e^{-2mt_Q^2})$ increases, indicating that the bound holds with higher confidence.

C Additional theoretical statements

C.1 Supporting propositions

Proposition C.1. *Let $f : \mathcal{X} \to \mathcal{Y}$ be an invertible univariate function, where $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. Let $C = [y_{\text{lo}}, y_{\text{hi}}] \subseteq \mathcal{Y}$ be a closed interval. Then for any $y \in \mathcal{Y}$, the following equivalence holds:*

$$y \in C \quad\Longleftrightarrow\quad f^{-1}(y) \in \{x \in \mathcal{X} : f(x) \in C\}.$$

Proof. Since f is an invertible univariate function, it must be strictly monotonic, either strictly increasing or decreasing.

Case 1: Suppose f is strictly increasing. Then f^{-1} is also strictly increasing.

$\Rightarrow$ If $y \in C = [y_{\text{lo}}, y_{\text{hi}}]$, then by monotonicity,

$$f^{-1}(y_{\text{lo}}) \leq f^{-1}(y) \leq f^{-1}(y_{\text{hi}}),$$

so $f^{-1}(y) \in [f^{-1}(y_{\text{lo}}), f^{-1}(y_{\text{hi}})]$. Since f is strictly increasing, this implies

$$[f^{-1}(y_{\text{lo}}), f^{-1}(y_{\text{hi}})] = \{x \in \mathcal{X} : f(x) \in C\}$$

and thus $f^{-1}(y) \in \{x \in \mathcal{X} : f(x) \in C\}$.

$\Leftarrow$ If $f^{-1}(y) \in \{x \in \mathcal{X} : f(x) \in C\}$, then equivalently we can derive $y \in C$.

Case 2: Suppose f is strictly decreasing. Then f^{-1} is also strictly decreasing.

$\Rightarrow$ If $y \in C = [y_{\text{lo}}, y_{\text{hi}}]$, then

$$f^{-1}(y_{\text{lo}}) \geq f^{-1}(y) \geq f^{-1}(y_{\text{hi}}),$$

so $f^{-1}(y) \in [f^{-1}(y_{\text{hi}}), f^{-1}(y_{\text{lo}})]$. Again, since f is decreasing,

$$[f^{-1}(y_{\text{hi}}), f^{-1}(y_{\text{lo}})] = \{x \in \mathcal{X} : f(x) \in C\},$$

which implies $f^{-1}(y) \in \{x \in \mathcal{X} : f(x) \in C\}$.

$\Leftarrow$ If $f^{-1}(y) \in \{x \in \mathcal{X} : f(x) \in C\}$, then again $y \in C$.

In either case, the equivalence holds. □

Proposition C.2. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a univariate function, where $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. Let $C \subseteq \mathcal{Y}$ be a closed interval. Then for $a \in \mathcal{X}$, it holds that:*

$$f(a) \in C \iff a \in \{x \in \mathcal{X} : f(x) \in C\}.$$

Proof. The statement is a direct consequence of the definition of the set $\{x \in \mathcal{X} : f(x) \in C\}$. By definition, a belongs to this set if and only if $a \in \mathcal{X}$ and $f(a) \in C$. Since $a \in \mathcal{X}$ is already assumed, the condition reduces to: $f(a) \in C \iff a \in \{x \in \mathcal{X} : f(x) \in C\}$. □

We visualize Proposition C.1 and Proposition C.2 in Figure 8.

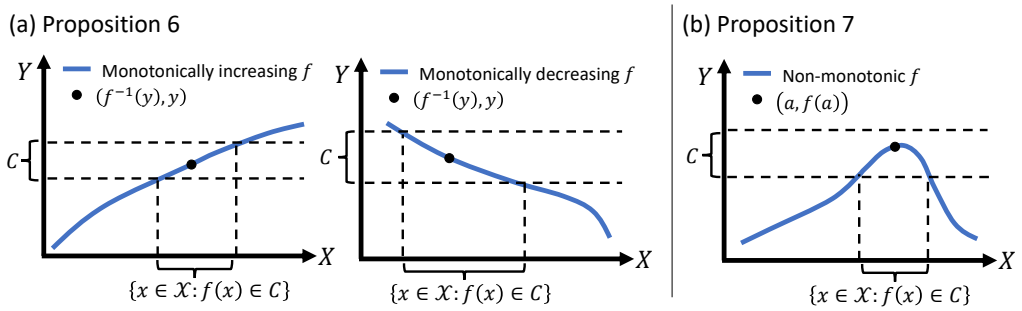


Figure 8: Characterization of preimage membership under **(a)** monotonic and **(b)** non-monotonic functions.

C.2 Proof of Theorem 3.2

Proof. Let μ_{XY} and ν_{XY} be probability measures on the metric space $(\mathcal{X} \times \mathcal{Y}, d_{\mathcal{XY}})$, where $d_{\mathcal{XY}}((x_1, y_1), (x_2, y_2)) := ||(d_{\mathcal{X}}(x_1, x_2), d_{\mathcal{Y}}(y_1, y_2))||_2$. Let $s : \mathcal{X} \times \mathcal{Y} \to \mathcal{V}$ be a measurable function such that $s(x, y) = v$. In metric space $(\mathcal{V}, d_{\mathcal{V}})$, denote μ_V the probability measure of $s(X, Y)$ for $(X, Y) \sim \mu_{XY}$. Also, let ν_V be the probability measure of $s(X, Y)$ for $(X, Y) \sim \nu_{XY}$. Denote $\Gamma_{V|x} = \Gamma(\mu_{V|x}, \nu_{V|x})$ and $\Gamma_{Y|x} = \Gamma(\mu_{Y|x}, \nu_{Y|x})$. By Theorem 1 in [8], we derive

$$\begin{aligned} W(\mu_{V|x}, \nu_{V|x}) &= \inf_{\gamma \in \Gamma_{V|x}} \int_{\mathcal{V} \times \mathcal{V}} d_{\mathcal{V}}(v_1, v_2) \, \mathrm{d}\gamma(v_1, v_2) \\ &= \inf_{\gamma \in \Gamma_{Y|x}} \int_{\mathcal{Y} \times \mathcal{Y}} d_{\mathcal{V}}(s(x, y_1), s(x, y_2)) \, \mathrm{d}\gamma(y_1, y_2). \end{aligned} \tag{33}$$

Consider $\gamma^* \in \Gamma_{Y|x}$ is the optimal transport plan for $W(\mu_{Y|x}, \nu_{Y|x})$. However, γ^* is not necessarily optimal for obtaining $W(\mu_{V|x}, \nu_{V|x})$ in Eq. (33), so we have

$$W(\mu_{V|x}, \nu_{V|x}) \leq \int_{\mathcal{Y}\times\mathcal{Y}} d_{\mathcal{V}}(s(x, y_1), s(x, y_2))\, \mathrm{d}\gamma^*(y_1, y_2). \tag{34}$$

Given that the function s is continuous with constant κ conditioned on x, we have $\frac{d_{\mathcal{V}}(s(x,y_1),s(x,y_2))}{d_{\mathcal{Y}}(y_1,y_2)} \leq \kappa, \forall x \in \mathcal{X}, y_1, y_2 \in \mathcal{Y}$, so the following inequality holds that

$$\int_{\mathcal{Y}\times\mathcal{Y}} d_{\mathcal{V}}(s(x, y_1), s(x, y_2))\, \mathrm{d}\gamma^*(y_1, y_2) \leq \int_{\mathcal{Y}\times\mathcal{Y}} \kappa \cdot d_{\mathcal{Y}}(y_1, y_2)\, \mathrm{d}\gamma^*(y_1, y_2) = \kappa \cdot W(\mu_{Y|x}, \nu_{Y|x}). \tag{35}$$

Finally, combining Eq. (34) and Eq. (35), we can conclude that

$$W(\mu_{V|x}, \nu_{V|x}) \leq \kappa \cdot W(\mu_{Y|x}, \nu_{Y|x}). \tag{36}$$

□

C.3 Proof of Theorem 3.3

Proof. Let γ^*_{XYXY} be the optimal transport plan satisfying the assumptions of Theorem 3.3, and let ψ be its diagonal transport measure.

Since $d_{\mathcal{XY}}$ is the 2-product metric,

$$d_{\mathcal{XY}}\big((x_1, y_1), (x_2, y_2)\big) \geq d_{\mathcal{Y}}(y_1, y_2).$$

By optimality of γ^*_{XYXY}, disintegrating the transport cost and restricting it to $x_1 = x_2$ gives

$$\begin{aligned} W(\mu_{XY}, \nu_{XY}) &\geq \int_{\mathcal{X}} \left[\int_{\mathcal{Y}^2} d_{\mathcal{Y}}(y_1, y_2)\, \mathrm{d}\gamma^*_{YY|x,x}(y_1, y_2) \right] \mathrm{d}\psi(x) \\ &\geq \int_{\mathcal{X}} W(\mu_{Y|x}, \nu_{Y|x})\, \mathrm{d}\psi(x) \\ &\geq c \int_{\mathcal{X}} W(\mu_{Y|x}, \nu_{Y|x})\, \mathrm{d}\nu_X(x). \end{aligned} \tag{37}$$

The second inequality follows because $\gamma^*_{YY|x,x}$ is an admissible coupling of $\mu_{Y|x}$ and $\nu_{Y|x}$ for ψ-almost every x. The last inequality follows from $\psi \geq c\nu_X$ and the nonnegativity of the Wasserstein distance.

Dividing by $c > 0$ and setting $\eta := 1/c$ yields

$$\int_{\mathcal{X}} W(\mu_{Y|x}, \nu_{Y|x})\, \mathrm{d}\nu_X(x) \leq \eta\, W(\mu_{XY}, \nu_{XY}),$$

which completes the proof. □

C.4 Proof of Theorem 5.1

Proof. For each $k \in \{1, ..., K\}$, denote $\gamma^k \in \Gamma(\mu_{XY}, \nu^k_{XY})$ the optimal transport plan realizing $W(\mu_{XY}, \nu^k_{XY})$ such that

$$W(\mu_{XY}, \nu^k_{XY}) = \int_{\mathcal{X}\times\mathcal{Y}} d_{\mathcal{XY}}(x, y) \mathrm{d}\gamma^k(x, y). \tag{38}$$

Given $\nu_{XY} = \sum_{k=1}^K \lambda_k \nu^k_{XY}$, let $\gamma^* = \sum_{k=1}^K \lambda_k \gamma^k$. Since the first marginal of γ^* is μ_{XY} and the second marginal of γ^* is $\sum_{k=1}^K \lambda_k \nu^k_{XY}$, it follows that $\gamma^* \in \Gamma(\mu_{XY}, \nu_{XY})$. However, γ^* is not necessarily an optimal transport plan for $W(\mu_{XY}, \nu_{XY})$, we conclude that

$$\begin{aligned} &W(\mu_{XY}, \nu_{XY}) = \inf_{\gamma\in\Gamma(\mu_{XY},\nu_{XY})} \int_{\mathcal{X}\times\mathcal{Y}} d_{\mathcal{XY}}(x, y) \mathrm{d}\gamma(x, y) \leq \int_{\mathcal{X}\times\mathcal{Y}} d_{\mathcal{XY}}(x, y) \mathrm{d}\gamma^*(x, y) \\ &= \sum\nolimits_{k=1}^K \lambda_k \int_{\mathcal{X}\times\mathcal{Y}} d_{\mathcal{XY}}(x, y) \mathrm{d}\gamma^k(x, y) = \sum\nolimits_{k=1}^K \lambda_k W(\mu_{XY}, \nu^k_{XY}). \end{aligned} \tag{39}$$

□

C.5 Proof of Theorem B.4

Proof. By the triangle inequality and symmetry of the Wasserstein distance,

$$\begin{aligned} W(\mu_X, \nu_X) &\le W(\mu_X, \widehat{\mu}_X) + W(\widehat{\mu}_X, \widehat{\nu}_X) + W(\widehat{\nu}_X, \nu_X), \\ W(\widehat{\mu}_X, \widehat{\nu}_X) &\le W(\widehat{\mu}_X, \mu_X) + W(\mu_X, \nu_X) + W(\nu_X, \widehat{\nu}_X). \end{aligned} \tag{40}$$

Rearranging these two inequalities gives

$$|W(\mu_X, \nu_X) - W(\widehat{\mu}_X, \widehat{\nu}_X)| \le W(\mu_X, \widehat{\mu}_X) + W(\nu_X, \widehat{\nu}_X).$$

Since $\sigma_\mu > d_W(\mu_X)$ and $\sigma_\nu > d_W(\nu_X)$, Theorem B.3 provides constants $\lambda_\mu, \lambda_\nu > 0$ such that

$$\mathbb{E}[W(\mu_X, \widehat{\mu}_X)] \le \lambda_\mu n^{-1/\sigma_\mu}, \qquad \mathbb{E}[W(\nu_X, \widehat{\nu}_X)] \le \lambda_\nu m^{-1/\sigma_\nu}.$$

Combining these expectation bounds with the concentration inequalities in Theorem B.3 yields, for any $t_\mu, t_\nu > 0$,

$$\begin{aligned} \Pr\Big(W(\mu_X, \widehat{\mu}_X) \le \lambda_\mu n^{-1/\sigma_\mu} + t_\mu\Big) &\ge 1 - e^{-2nt_\mu^2}, \\ \Pr\Big(W(\nu_X, \widehat{\nu}_X) \le \lambda_\nu m^{-1/\sigma_\nu} + t_\nu\Big) &\ge 1 - e^{-2mt_\nu^2}. \end{aligned} \tag{41}$$

Each event depends only on its corresponding sample. Because the two samples are independent, both events hold simultaneously with probability at least

$$\left(1 - e^{-2nt_\mu^2}\right)\left(1 - e^{-2mt_\nu^2}\right).$$

On their intersection, the preceding triangle-inequality bound implies

$$|W(\mu_X, \nu_X) - W(\widehat{\mu}_X, \widehat{\nu}_X)| \le \lambda_\mu n^{-1/\sigma_\mu} + \lambda_\nu m^{-1/\sigma_\nu} + t_\mu + t_\nu. \tag{42}$$

This proves the theorem. □

D Demonstration of implicit dependency

We demonstrate that $f_{\theta_Y}^{-1}(\overline{Y}_{n+1})$ implicitly depends on X_{n+1} through the composition $\phi \circ f_{\theta_X}$, where $\phi : \mathcal{X} \to \mathcal{Y}$ denotes the possibly stochastic labeling mechanism under P_{XY}. Consider a BNF f_θ is optimized by Wasserstein distance minimization in Eq. (18) such that $f_{\theta\#}Q_{XY} = P_{XY}$. Therefore, for a test sample $(X_{n+1}, Y_{n+1}) = (x, y) \sim Q_{XY}$, it holds that

$$(\overline{X}_{n+1}, \overline{Y}_{n+1}) = f_\theta(x, y) = (f_{\theta_X}(x), f_{\theta_Y}(y)) = (\bar{x}, \bar{y}) \sim P_{XY}.$$

As a result, $\mathcal{C}_{\text{C}}(\overline{X}_{n+1})$ satisfies the conditional coverage guarantee under P_{XY}. Moreover, since $\bar{y} = \phi(\bar{x})$, we obtain

$$f_{\theta_Y}^{-1}(\bar{y}) = f_{\theta_Y}^{-1}(\phi(\bar{x})) = f_{\theta_Y}^{-1}(\phi(f_{\theta_X}(x))),$$

which shows that the inverse transformation $f_{\theta_Y}^{-1}(\bar{y})$ used to construct $\mathcal{C}_{\text{BNF}}(X_{n+1})$ inherently captures the dependency on $X_{n+1} = x$.

We present an example to illustrate the dependency. Denote $\mathcal{U}$ and $\mathcal{N}$ uniform and Gaussian distributions, respectively. To introduce a distribution shift between P_{XY} and Q_{XY}, let

$$\begin{aligned} P_X &= \mathcal{U}(0, 1), P_{Y|X} = \mathcal{N}(-0.5X, -0.3X^2 + 0.3X); \\ Q_X &= \mathcal{U}(0, 0.8), Q_{Y|X} = \mathcal{N}(0.25X, -0.24X^2 + 0.24X). \end{aligned}$$

Figure 9 shows how the inverse transformation $f_{\theta_Y}^{-1}$ preserves the conditional guarantee from $\mathcal{C}_{\text{C}}(\overline{X}_{n+1})$ to $\mathcal{C}_{\text{BNF}}(X_{n+1})$ through the implicit dependency on $X_{n+1} = x$.

We admit the factorized architecture of BNF relies on a mild structural assumption of P_{XY} and Q_{XY} to realize the exact alignment in Eq. (14).

Let labeling processes be $Y = \psi(X) + \psi_{\text{rand}}$ on Q_{XY} and $Y = \phi(X) + \phi_{\text{rand}}$ on P_{XY}, where ψ_{rand} and ϕ_{rand} are stochastic parts. Eq. (14) implies

$$f_{\theta_Y}(\psi(X) + \psi_{\text{rand}}) = \phi(f_{\theta_X}(X)) + \phi_{\text{rand}}. \tag{43}$$

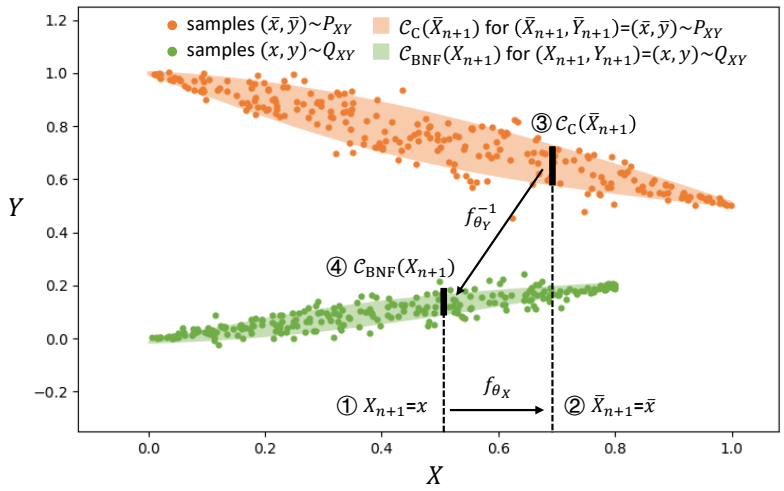


Figure 9: Preserving conditional coverage via implicit dependency on test input. The circled numbers indicate the sequential steps to obtain the corresponding values or prediction sets.

We implement f_{θ_Y} as an augmented Real NVP that includes tanh activations over Y. Hence, f_{θ_Y} is not affine with respect to $\psi(X)+\psi_{\text{rand}}$, so in general

$$f_{\theta_Y}(\psi(X)+\psi_{\text{rand}}) \neq f_{\theta_Y}(\psi(X)) + f_{\theta_Y}(\psi_{\text{rand}}). \tag{44}$$

Consequently, we cannot split Eq. 43 by $f_{\theta_Y}(\psi(X)) = \phi(f_{\theta_X}(X))$ and $f_{\theta_Y}(\psi_{\text{rand}}) = \phi_{\text{rand}}$. Accordingly, the deterministic-stochastic decomposition of Eq. 43 can be naturally expressed via expectation:

$$\phi(f_{\theta_X}) = \mathbb{E}_{\psi_{\text{rand}}}[f_{\theta_Y}(\psi(X)+\psi_{\text{rand}})], \tag{45}$$

$$\phi_{\text{rand}} = f_{\theta_Y}(\psi(X)+\psi_{\text{rand}}) - \mathbb{E}_{\psi_{\text{rand}}}[f_{\theta_Y}(\psi(X)+\psi_{\text{rand}})]. \tag{46}$$

This leads to a structural assumption: for exact alignment to hold, both ϕ and ϕ_{rand} should be generally feature-dependent, unless ψ_{rand} is independent noise and ψ is constant with respect to X.

The experimental result in Figure3 shows that the augmented factorized architecture is sufficiently expressive to capture practical distribution shifts, where both deterministic and stochastic labeling parts on P and Q are typically feature-dependent, satisfying the assumption above.

A possible way to eliminate this assumption is to condition the transformation of Y on the feature X. However, as discussed in Section 7.1, this approach introduces practical challenges, most notably exacerbating the curse of dimensionality.

E Comparison between normalizing flow techniques

The monotonicity of the univariate f_{θ_Y} allows us to take advantage of Proposition C.1 to inversely transform $\mathcal{C}_{\mathrm{A}}(\overline{X}_{n+1})$ via Eq. (21). However, the monotonicity also limits the flexibility of f_{θ_Y}, restricting the class of distributions it can model. Here, we briefly introduce several normalizing flow techniques designed for one-dimensional transformations that often struggle to map complex distributions effectively, thereby motivating the design of Augmented BNF in Section 4.2. For a more comprehensive overview of normalizing flows, we refer to the survey by [25].

We begin with planar flow, a fundamental transformation that expands or contracts the input space along specific directions [46]. A planar flow is achieved by applying a linear transformation followed by a nonlinear activation, which dictates how the data is warped. To enhance expressiveness, normalizing flows are typically constructed as compositions of multiple sub-flows. We implement a BNF where each branch applies a sequence of 16 planar flows. LeakyReLU is used as the nonlinear activation function to preserve invertibility throughout the transformation.

Residual flow is built using residual connections [47]. The output of a residual connection is the sum of the original input and a transformation generated by a neural network. For these residual connections to be invertible, the transformation must have a Lipschitz constant less than 1, ensuring

that the transformation does not distort the data too much. We also construct a BNF where each branch consists of 16 residual connections. The neural network within each residual connection has an architecture consisting of an input layer, two hidden layers with 128 units each, and an output layer matching the input dimension.

Both planar flow and residual flow are capable of transforming one-dimensional data. In addition, autoregressive flow [48, 26] offers an alternative approach by modeling each transformation step as conditioned on the preceding ones, meaning the transformation of each sample value explicitly depends on the values that came before it. This sequential dependency enables more flexible and expressive density estimation, particularly in one-dimensional settings. However, because BNF requires deterministic transformations that are independent of input ordering, autoregressive flow is not suitable for our approach.

We illustrate the performance of BNFs constructed using planar flow and residual flow in Figure 10 and compare them against the Augmented BNF, which is implemented using a standard coupling normalizing flow, Real NVP [28]. Detailed specifications for the Augmented BNF are provided in Appendix F. The results show that BNFs using univariate f_{θ_Y} struggle to transform complex distributions effectively, resulting in lower WSC and higher WSCG.

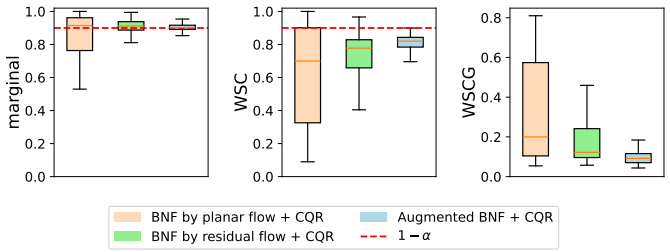


Figure 10: Marginal coverage, WSC, and WSCG of BNFs constructed with planar and residual flows, compared with Augmented BNF at confidence level $1 - \alpha = 0.9$.

F Architecture and Computational Cost of Augmented BNF

Augmented BNF uses coupling flows to learn the transformation and numerical inversion to construct prediction sets. This appendix describes the architecture and reports the computational costs of training and inference.

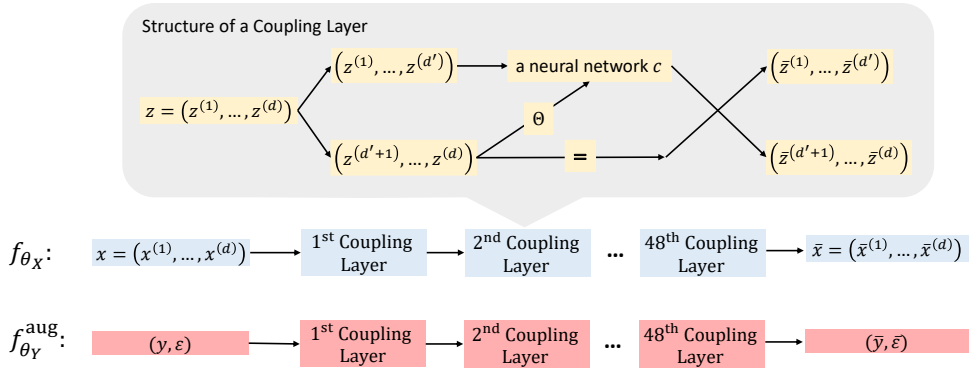


Figure 11: Illustration of the coupling layer structure and the overall composition of Augmented BNF.

F.1 Architecture via Coupling Flows

Both branches of Augmented BNF operate on multi-dimensional data, enabling the use of coupling flows—a technique for modeling complex high-dimensional distributions. A coupling flow usually

consists of multiple coupling layers. In a coupling layer, the input is partitioned into two parts. One part remains unchanged during the transformation, while the other is modified using a neural network c, whose parameter Θ depends on the unchanged part. This setup ensures invertibility and allows for flexible, learnable transformations. Afterward, a permutation step is applied for higher expressiveness. In our implementation, each branch of Augmented BNF consists of a sequence of 48 coupling layers based on Real NVP [28], allowing the entire input to be progressively transformed. The neural network c within each coupling layer follows a symmetric architecture with hidden layers of sizes 64, 128, 256, 128, and 64, mapping from the input dimension to the output dimension. Figure 11 illustrates the structure of a coupling layer, using a random variable $Z \in \mathbb{R}^d$ with a realization z, and shows how both branches are constructed by stacking multiple coupling layers. The normalized Gaussian noise $\bar{\varepsilon}$ is discarded after the transformation.

F.2 Computational Cost

Training. Augmented BNF introduces an additional flow training stage. The flow is trained once per experimental trial and reused across test mixtures. Table 2 reports its training time on an NVIDIA RTX 3090, excluding the training of the prediction models.

Table 2: Additional training time of Augmented BNF per experimental trial on an NVIDIA RTX 3090.

Setting	Training time
Bike Rental	8 min 50 s
PTS	12 min 30 s
Traffic	8 min 30 s
U.S. ILI	5 min 40 s
Fair Med	8 min 40 s

Inference. Unlike the original BNF, Augmented BNF does not employ a univariate monotonic transformation for the Y-branch. As a result, we cannot directly apply Eq. (20) to construct the prediction set. Instead, Augmented BNF relies on Eq. (24) to generate $\mathcal{C}^{\text{aug}}_{\text{BNF}}(X_{n+1})$, which frames the construction as a root-finding problem. Specifically, let the interval endpoints of the calibrated set $\mathcal{C}_{\text{C}}(\overline{X}_{n+1})$ be denoted by y_{lo} and y_{hi}. Then, Eq. (24) requires solving a root-finding problem to identify the pre-images of these endpoints under the learned transformation. In particular, we need to find the values of y that satisfy the following equations

$$f^{\text{aug}}_{\theta_Y}(y; \varepsilon_{n+1}) = y_{\text{lo}}; \quad f^{\text{aug}}_{\theta_Y}(y; \varepsilon_{n+1}) = y_{\text{hi}}. \tag{47}$$

In practice, we construct a prediction set by evaluating a coarse grid of 100 candidate y values and checking, via the forward transformation, whether the corresponding $\bar{y}$ falls inside $\mathcal{C}_{\text{C}}(\overline{X}_{n+1})$. Values that satisfy this condition are included in $\mathcal{C}^{\text{aug}}_{\text{BNF}}(X_{n+1})$, with the endpoints defining the prediction set boundary. In practice, 100 candidates are sufficient to accurately capture the boundaries, and adding more candidates has little effect on the set size. The procedure is computationally efficient, adding only 0.03 seconds per test sample compared to standard CQR on an RTX 3090.

For a lightweight evaluation of prediction-set size, we use 10 test sets per trial, totaling 100 test sets across 10 independent trials for each dataset and experimental configuration. We average prediction-set sizes over samples within each test set, then over test sets and trials. Coverage metrics are evaluated on all 100 test sets per trial.

G Conformalized quantile regression

The conditional guarantee in Eq. (3) is not practically achievable using finite calibration samples without regularity assumptions, such as Lipschitz continuity of $P_{Y|x}$ density [18]. Hence, approximations of the conditional guarantee are extensively developed. Mondrian CP ensures $1-\alpha$ coverage conditioned over input subspaces [49]. Some methods estimate the conformal score distribution conditioned on specific test input x, for example, by weighting each V_i based on the proximity of X_i to x [50, 51, 52]. Conformal training embeds a size-based loss in the training of the model h [53, 54, 55]. Besides, advanced score functions are developed to facilitate conditional coverage in

regression [30, 56]. Generative models also show promise for enhancing adaptiveness, especially for multivariate output [57, 58, 59, 60].

In this work, we apply Conformalized quantile regression (CQR) [30] to approximate the conditional guarantee in Eq. (3). CQR first trains two regression models with pinball loss at levels $1-\alpha/2$ and $\alpha/2$, respectively, then calibrates the resulting intervals using residuals on a separate calibration set.

For clarity, we introduce CQR in the context of sample normalization and multi-source domain generalization. For a regression model h, the pinball loss [61] at quantile level α for sample (x, y) is defined as

$$l_\alpha(h(x), y) = \begin{cases} \alpha\,(y - h(x)) & \text{if } y - h(x) > 0, \\ (1-\alpha)\,(h(x) - y) & \text{otherwise.} \end{cases} \tag{48}$$

The models h_{hi} and h_{lo} are trained by optimizing the pinball loss in Eq (48) at quantile levels $1-\alpha/2$ and $\alpha/2$, respectively. For calibration instances $\{(X_i, Y_i)\}_{i=1}^n$ drawn from P_{XY}, conformal scores are defined as

$$V_i = \max\left\{h_{\text{lo}}(X_i) - Y_i, Y_i - h_{\text{hi}}(X_i)\right\}, \text{for } i = 1, ..., n. \tag{49}$$

Let τ be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of $\{V_i\}_{i=1}^n$. If a test sample $(X_{n+1}, Y_{n+1}) \sim Q_{XY}$ is normalized to $(\overline{X}_{n+1}, \overline{Y}_{n+1}) \sim P_{XY}$, we construct an adaptive prediction set

$$\mathcal{C}_{\text{CQR}}(\overline{X}_{n+1}) = \left[h_{\text{lo}}(\overline{X}_{n+1}) - \tau, h_{\text{hi}}(\overline{X}_{n+1}) + \tau\right]. \tag{50}$$

Here, h_{lo} and h_{hi} predict the likely lower and upper ends, while τ adjusts the set based on how well the predictions fit the calibration data. As proved in [30], $\mathcal{C}_{\text{CQR}}$ can empirically approximate the conditional coverage guarantee described in Eq. (19). Extensions of CQR are explored in [62, 63], which modified the score function in Eq. (49) for higher adaptiveness.

H A brief review of the Sinkhorn algorithm

As we introduced in Definition B.1, the Wasserstein distance between two empirical distributions $\widehat{\mu}_X$ and $\widehat{\nu}_X$ with $p = 1$ is given by

$$W(\widehat{\mu}_X, \widehat{\nu}_X) = \min_{\gamma \in \Gamma(\widehat{\mu}_X, \widehat{\nu}_X)} \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij}\, C_{ij}.$$

where $C \in \mathbb{R}^{n \times m}$ is the cost matrix with entries $C_{ij} = d_{\mathcal{X}}(x_i, x'_j)$, and $\Gamma(\widehat{\mu}_X, \widehat{\nu}_X)$ is the set of joint distributions $\gamma \in \mathbb{R}_+^{n \times m}$ with marginals $\widehat{\mu}_X$ and $\widehat{\nu}_X$.

To make this optimization problem more tractable, the Sinkhorn algorithm [32] introduces an entropic regularization term:

$$W^\beta(\widehat{\mu}_X, \widehat{\nu}_X) = \min_{\gamma \in \Gamma(\widehat{\mu}_X, \widehat{\nu}_X)} \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij}\, C_{ij} + \beta \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} \log \gamma_{ij},$$

where $\beta > 0$ controls the strength of the regularization.

This regularized objective is strictly convex and can be efficiently minimized via iterative matrix scaling. Let $K = \exp(-C/\beta)$ be the Gibbs kernel. The scaling vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ are initialized to all ones and updated via

$$u \leftarrow \frac{1/n}{Kv}, \quad v \leftarrow \frac{1/m}{K^\top u},$$

where divisions are element-wise. Once converged with small changes in u and v, the optimal transport plan takes the form

$$\gamma^* = \operatorname{diag}(u)\, K \operatorname{diag}(v).$$

This approach yields a differentiable approximation to the true Wasserstein distance, enabling its integration into gradient-based optimization pipelines. We refer to [32, 33, 64] for more detailed studies about the Sinkhorn algorithm.

In our implementation, the flow-alignment loss is computed using the debiased Sinkhorn divergence provided by GeomLoss [34], with $p = 2$ and quadratic ground cost $c(z, z') = \frac{1}{2}\|z - z'\|_2^2$. While W denotes the 1-Wasserstein distance in our theoretical analysis, we use this quadratic-cost Sinkhorn divergence as a practical training surrogate for distributional alignment. The theoretical bounds remain expressed in terms of W_1; we do not directly substitute the empirical Sinkhorn loss into these bounds.

I Introduction to baselines

Figure 12 highlights the distinctions between the baseline methods and the proposed approach. SCP constructs prediction sets of fixed size and ensures only marginal coverage under i.i.d. assumptions, rendering it ineffective under joint distribution shifts. IW-CP addresses only covariate shift and causes its prediction intervals to contract in the example, because test features are distributed in regions where calibration data is concentrated. WC-CP accounts for worst-case distribution shifts, expanding prediction sets until $1-\alpha$ marginal coverage is achieved on the test data, which can be inefficient. WR-CP improves upon this by regularizing the base predictive model through minimizing the Wasserstein distance between calibration and test conformal scores, producing more compact prediction sets while maintaining robust marginal coverage. All of these methods, however, focus exclusively on marginal coverage. CQR, a representative conditional conformal prediction method, fails to handle distributional shifts. In contrast, the Augmented BNF transformation model learns an invertible mapping between calibration and test data, enabling robust conditional coverage even under non-i.i.d. conditions.

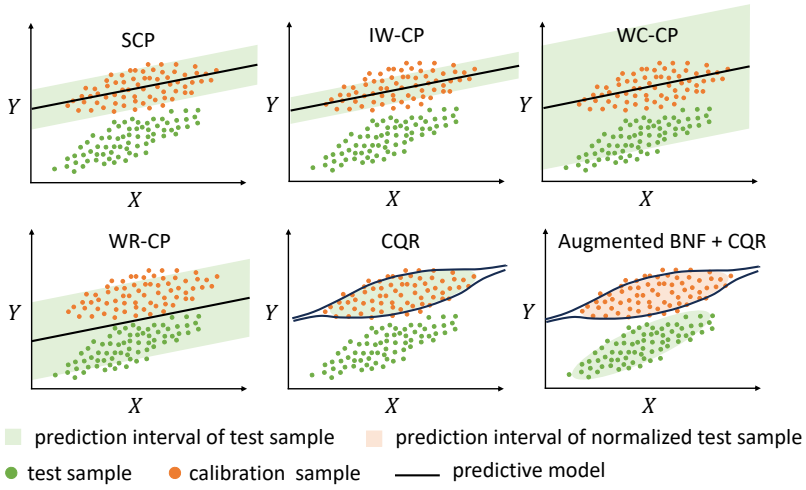


Figure 12: Comparison between baselines and the proposed method via a toy example. Augmented BNF effectively transforms the prediction intervals from the calibration distribution to the test distribution.

J Data preparation for multi-source domain generalization

We introduce the data preparation procedure shared across all datasets. We set $K = 3$, partitioning each dataset into three source subsets with distinct distributions. We take the calibration distribution to be the equally weighted mixture $P_{XY} := \frac{1}{K}\sum_{k=1}^{K} D_{XY}^{k}$.

For each dataset, we conduct 10 independent sampling trials. In each trial, we sample equally sized training sets S_{D^k} from their respective sources without replacement. Their union $S_P^{\text{align}} := \bigcup_{k=1}^{K} S_{D^k}$ provides an empirical approximation of P_{XY} and is used for training Augmented BNF. From the remaining observations, we construct S_P^{cal} by sampling equal numbers of observations from the K sources and reserve it exclusively for conformal calibration. Finally, we construct the random-mixture test sets from observations excluded from all training and calibration sets. Experimental results are aggregated over the 10 trials for each dataset.

We also leverage a toy example from [8] to demonstrate joint distribution shift under multi-source domain generalization in Figure 13.

J.1 Synthetic distribution shifts

The Physicochemical Properties of Protein Tertiary Structure (PTS) dataset [9] contains 45,730 instances, with the target variable being the protein decoy size. It includes nine features: surface area,

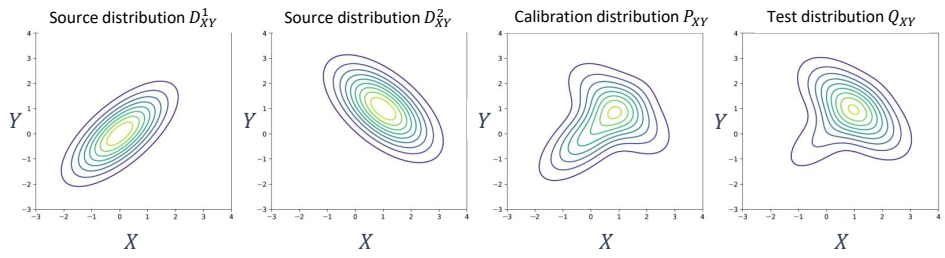


Figure 13: **Multi-source domain generalization.** The test distribution Q_{XY} is a random mixture of source distributions, while the calibration distribution P_{XY} is a fixed known distribution. As a result, a distribution shift occurs since $P_{XY} \neq Q_{XY}$.

non-polar exposed area, fractional area of exposed non-polar residue, fractional area of exposed non-polar part, molecular mass weighted exposed area, average deviation, Euclidean distance, secondary structure penalty, and spatial distribution constraints. Raw data is split into three subsets based on the distribution of the secondary structure penalty, thereby introducing distribution shifts among the subsets. We also use the PTS dataset to perform ablation studies on the approximation ability (with varying sample sizes) and generalization performance (with different numbers of source domains) of Augmented BNF.

J.2 Natural distribution shifts reflecting real-world challenges

Generalized sales prediction over time-series is crucial for risk-averse business decision-making [65]. Moreover, sales data typically exhibit strong periodic patterns, such as seasonal or weekly fluctuations. Thus, effectively utilizing data from each sub-period to model a robust and generalized sales pattern is critical for achieving reliable forecasts. This requires models not only to capture short-term variations but also to generalize across different temporal domains, where distribution shifts may occur naturally due to changes in consumer behavior, external events, or market conditions. We consider the Bike Rental dataset [10] to reflect this challenge. The dataset records hourly and daily rental counts from the Capital Bikeshare system during 2011 and 2012, along with associated weather and seasonal information. We partition the data based on rental hours into three time intervals: [0,8] (midnight), [9,16] (daytime), and [17,23] (evening). For prediction, we select continuous features including temperature, feeling temperature, humidity, and wind speed. The target variable is the count of rental bikes.

Traffic speed prediction with mismatched data focuses on transferring models trained on source distributions (e.g., traffic patterns on regular days and at major intersections) to test distributions exhibiting different characteristics (e.g., traffic patterns on special days and at minor intersections). For example, recent work has proposed traffic-law-informed models based on reaction-diffusion equations to provide generalized speed predictions [66]. Nevertheless, enhancing the reliability of uncertainty quantification under such distribution shifts remains a significant challenge. The Seattle-Loop dataset contains traffic volume and speed data collected in Seattle throughout 2015, recorded by sensors at 5-minute intervals [11]. PEMSD4 includes traffic data from 29 roads in San Francisco collected between January and February 2018, while PEMSD8 covers 8 roads in San Bernardino from July to August 2016 [37]. The task is to predict traffic speed at the next time step based on current speed and volume measurements. With $K = 3$, we select one representative intersection from each dataset. Due to varying local traffic patterns, natural distribution shifts arise among the three locations. Our goal is to achieve strong generalization across these locations, ensuring robust predictions on any test sites where traffic patterns resemble a random mixture of the three selected intersections.

Fair medical decision-making for patients from different hospitals is essential for ensuring equitable healthcare outcomes. Variations in patient demographics, medical imaging scanners, laboratory equipment, and clinical practices across hospitals can lead to distribution shifts in the data. This phenomenon is commonly referred to as the multi-center issue [67]. Addressing this challenge is essential for building predictive models that remain accurate and fair across diverse healthcare institutions [68]. To validate the effectiveness of the proposed method in this task, we collect patient

data from a collaborating hospital. Additionally, we use the MIMIC-IV [12] and eICU [13] datasets to simulate data from two other hospitals. The goal is to fairly predict patients' ICU stay times based on their Apache scores and blood urea nitrogen (BUN) levels, ensuring reliable performance regardless of which center a patient originates from. We consider fair medical prediction to be achieved across the three data sources if the model exhibits comparable performance on random mixtures of the sources.

Robust epidemic modeling across pandemic phases can facilitate timely public health responses and resource planning. The U.S. Centers for Disease Control and Prevention (CDC) categorizes an epidemic period into three main phases: initiation, acceleration, and deceleration [69]. Each of these phases exhibits distinct epidemiological characteristics, which lead to natural distribution shifts. Traditional forecasting methods typically rely on Susceptible-Infectious-Recovered (SIR) models [70, 71, 72] to predict the number of recently infected patients, aiming for robustness across the different pandemic phases. We demonstrate the application of the proposed method using the U.S. Influenza-like Illness (ILI) dataset [14], which contains weekly reports from the CDC on the number of ILI patients. The objective is to predict new infections for the upcoming week using both the weekly increase of infected patients and the cumulative infections for the year. The raw data is divided into three subsets based on the corresponding pandemic phases. We consider a forecasting model to be robust if its predictions remain reliable on random mixtures of data from the three phases.

We further apply t-SNE [40] to map the samples from each source into two dimensions, as shown in Figure 14. The visualization reveals clear distributional shifts between most sources. However, for some cases, such as the second and third sources in the Bike Rental and Fair Med setups, the distributions appear more similar. This slight overlap is not the result of manually creating similar data but arises naturally from the datasets themselves, which are collected from real-world scenarios.

Table 3 summarizes the datasets and source configurations. For Traffic and Fair Med, the three datasets constitute the three sources; for the remaining settings, the sources are constructed by partitioning one dataset.

Table 3: Summary of the nine datasets used in five experimental settings. d denotes the feature dimension, and $|S_{D^k}|$ is the number of training observations per source in each trial. Each setting uses $K = 3$ sources.

Setting	Dataset	d	$\lvert S_{D^k}\rvert$	Source construction
Bike Rental	Bike Rental	4	2,800	Three rental-hour intervals: $[0, 8]$, $[9, 16]$, and $[17, 23]$.
PTS	Physicochemical Properties of Protein Tertiary Structure	9	7,500	Three subsets defined by secondary structure penalty.
Traffic	Seattle-Loop	3	2,800	One selected intersection from Seattle-Loop.
	PEMSD4	3	2,800	One selected intersection from PEMSD4.
	PEMSD8	3	2,800	One selected intersection from PEMSD8.
U.S. ILI	U.S. Influenza-like Illness	2	870	Three epidemic phases: initiation, acceleration, and deceleration.
Fair Med	MIMIC-IV	2	3,000	One source formed from MIMIC-IV records.
	eICU	2	3,000	One source formed from eICU records.
	Collaborating hospital	2	3,000	One source formed from the collaborating hospital's records.

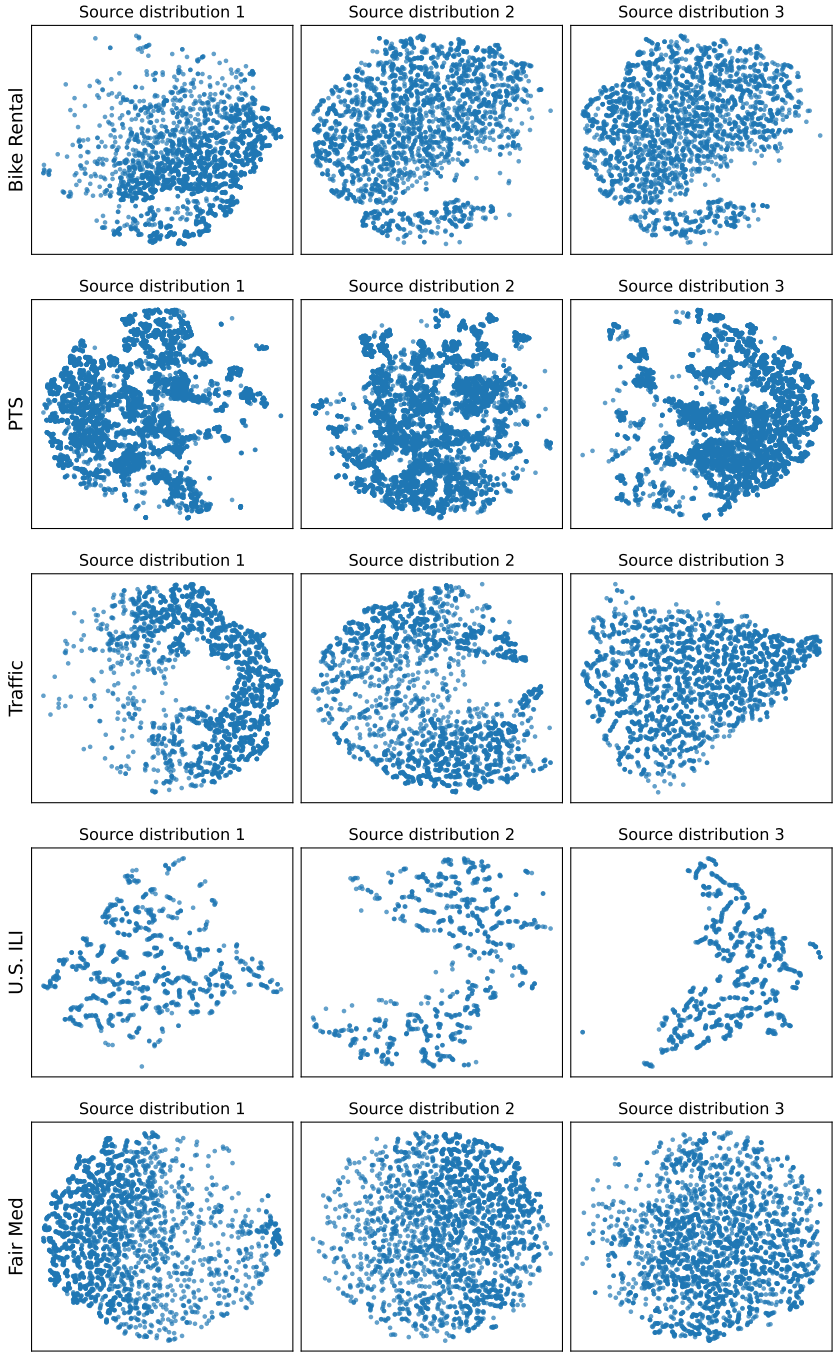


Figure 14: Empirical data distributions of source domains after applying t-SNE.

K Worst-slice coverage (WSC)

Worst-slice coverage (WSC) [38] quantifies the minimum empirical coverage over any slab $\mathcal{S} \subseteq \mathcal{X}$ that contains at least 10% of the test samples in $\mathcal{S}_Q$. Specifically, for any CP methods that produce a prediction set $C(x)$ given an input x, WSC is defined by

$$\text{WSC} = \inf_{\mathcal{S} \subseteq \mathcal{X}} \Pr(y \in C(x) | x \in \mathcal{S}), \text{s.t.} \Pr(x \in \mathcal{S} | (x, y) \in \mathcal{S}_Q) \geq 0.1. \tag{51}$$

Nevertheless, WSC only evaluates the infimum slice coverage and therefore fails to penalize over-coverage. As also noted in [73], ensuring a high worst-case slice does not guarantee good conditional coverage across $\mathcal{X}$, particularly when different regions exhibit excessive coverage.

These limitations motivate our introduction of WSCG in Eq. (27), which penalizes both under- and over-coverage from the target level $1-\alpha$. As a result, WSCG provides a more comprehensive and balanced assessment of conditional coverage robustness.

We approximate WSCG using slabs $S(v,a,b)=\{x : a \le v^\top x \le b\}$. We sample 1,000 Gaussian directions normalized to unit length and use 50 empirical projection quantiles as candidate endpoints for each direction. Among all endpoint pairs defining slabs containing at least 10% of the test observations, we report the maximum absolute deviation of empirical coverage from $1-\alpha$ as WSCG. On Bike Rental, minimum slice masses of 5%, 10%, and 20% yield similar mean WSCG across confidence levels $1-\alpha \in [0.1, 0.9]$, suggesting limited sensitivity to this threshold over the evaluated range.

L Generalization performance of Augmented BNF

L.1 Various numbers of source domains

To explore the generalization ability of Augmented BNF under varying numbers of source domains, we modified the sampling procedure in Appendix J by changing $K \in \{2,3,4,6,8,12\}$. For each value of K, we generated 10 independent trials using the PTS dataset to account for sampling variability. Augmented BNF combined with CQR was applied to each trial across confidence levels $1-\alpha \in [0.1, 0.9]$, enabling a comprehensive evaluation. Figure 15 shows that increasing the number of source domains does not significantly degrade conditional coverage robustness, suggesting that Augmented BNF generalizes well on the PTS dataset.

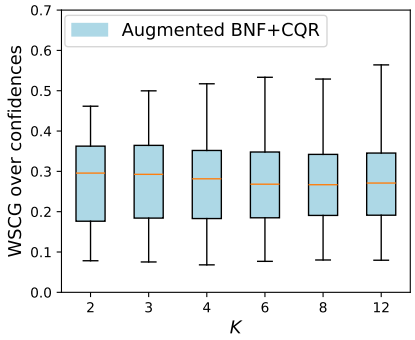


Figure 15: Generalization performance with different numbers of source domains.

L.2 Different sample sizes

Generative models may struggle to approximate underlying distributions when data are limited, especially in high dimensions [74, 75]. To assess how sample size affects the performance of Augmented BNF, we vary the number of samples in $\mathcal{S}_{D^k}$ and perform 10 trials for each setting on the PTS dataset. For each trial, we apply Augmented BNF+CQR across $1-\alpha$ from 0.1 to 0.9 and compute the WSCG over all confidence levels. As shown in Figure 16, WSCG degrades clearly when the size drops below 750. With 11-dimensional data on PTS, it gives a sample-to-dimension ratio of 68, demonstrating that our method is reasonably data-efficient.

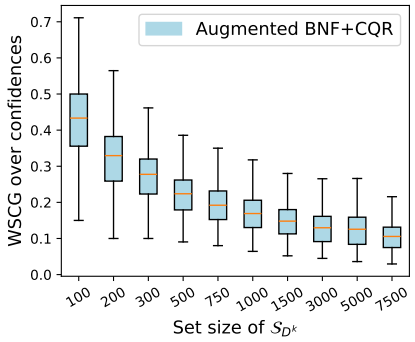


Figure 16: Impact of data availability on Augmented BNF approximation ability.

M Coverage lower bounds under imperfect transformation

M.1 Marginal coverage lower bound

We establish a marginal coverage lower bound by quantifying the alignment between the calibration conformal scores and those obtained from the transformed test distribution.

Marginal coverage gap can be defined as the discrepancy between the CDFs of P_V and Q_V at the calibration quantile τ [8]. After applying the transformation f_θ^{aug}, the test distribution Q_{XY} is mapped

to $\overline{Q}_{XY} := {f_\theta^{\text{aug}}}_{\#} Q_{XY}$, yielding the conformal score distribution $\overline{Q}_V := s_{\#}\overline{Q}_{XY}$, where s denotes the score function. The residual marginal coverage gap after transformation is therefore

$$|F_V^P(\tau) - F_V^{\overline{Q}}(\tau)|, \tag{52}$$

with F denoting the CDF. In this subsection, CDF expressions at the random threshold τ denote the corresponding coverage probabilities averaged over calibration sampling. We assume that the calibrated prediction set has marginal coverage at least $1-\alpha$ under P_{XY}, i.e., $F_V^P(\tau) \geq 1-\alpha$.

This leads to the following lower bound on the marginal coverage of prediction sets produced by the Augmented BNF transformation model:

$$\Pr\big(Y_{n+1} \in \mathcal{C}_{\text{BNF}}^{\text{aug}}(X_{n+1})\big) \geq 1 - \alpha - |F_V^P(\tau) - F_V^{\overline{Q}}(\tau)|. \tag{53}$$

Within the multi-source domain generalization (MSDG) framework, the test distribution Q_{XY} is assumed to be a random mixture of source distributions $\{D_{XY}^k\}_{k=1}^K$. Denoting $\overline{D}_{XY}^k := {f_\theta^{\text{aug}}}_{\#} D_{XY}^k$ and $\overline{D}_V^k := s_{\#}\overline{D}_{XY}^k$, we can bound the marginal coverage gap as

$$|F_V^P(\tau) - F_V^{\overline{Q}}(\tau)| \leq \sup_{k\in\{1,\ldots,K\}} |F_V^P(\tau) - F_V^{\overline{D}^k}(\tau)|. \tag{54}$$

Consequently, we obtain the final marginal coverage lower bound under MSDG as

$$\Pr\big(Y_{n+1} \in \mathcal{C}_{\text{BNF}}^{\text{aug}}(X_{n+1})\big) \geq 1 - \alpha - \sup_{k\in\{1,\ldots,K\}} |F_V^P(\tau) - F_V^{\overline{D}^k}(\tau)|. \tag{55}$$

For validation, we compare the theoretical bound with the empirical marginal coverage observed across randomly sampled test distributions. The results, presented in Figure 17, show that the empirical coverage for most test distributions exceeds the proposed lower bound, thereby confirming the validity of Eq. (55). The closeness between $1-\alpha$ and the bound further indicates that our method effectively aligns the calibration and test distributions.

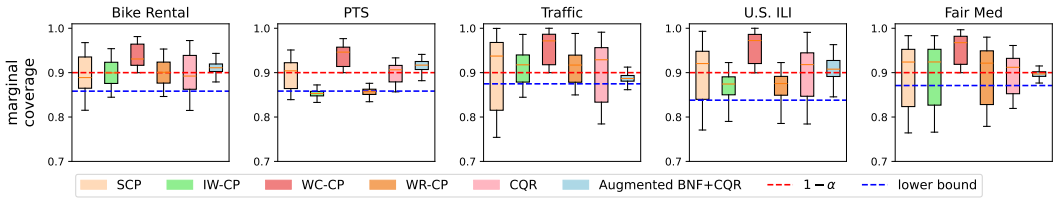


Figure 17: Marginal coverage achieved by $\mathcal{C}_{\text{BNF}}^{\text{aug}}(X_{n+1})$ compared with the proposed marginal lower bound.

M.2 Conditional coverage lower bound

We establish a conditional coverage lower bound that accounts for the imperfect alignment between calibration and test data induced by Augmented BNF.

Even under i.i.d. sampling, the marginal coverage guarantee of CQR does not imply $1-\alpha$ conditional coverage at every input. To account for this approximation error, we introduce $\alpha_{\text{i.i.d.}} \geq 0$ as a distribution-dependent upper bound on how far the conditional coverage of the fitted prediction set C_{CQR} can fall below $1-\alpha$, uniformly over x. Thus, for $(X_{n+1}, Y_{n+1}) \sim P_{XY}$,

$$\Pr(Y_{n+1} \in \mathcal{C}_{\text{CQR}}(X_{n+1}) | X_{n+1} = x) \geq 1 - \alpha - \alpha_{\text{i.i.d.}}. \tag{56}$$

This gap is an intrinsic limitation of existing conditional CP approaches.

Secondly, under distribution shift, the conditional coverage is further affected by the difference between the calibration and test score distributions. In this appendix, CCG and ICG are evaluated using the CQR score in Eq. (49) and its calibrated threshold τ, replacing the population conditional threshold $\tau(x)$. For $(X_{n+1}, Y_{n+1}) \sim Q_{XY}$, the CDF comparison used in Eq. (6), together with Eq. (56), gives

$$\Pr\left(Y_{n+1} \in \mathcal{C}_{\text{CQR}}(X_{n+1}) | X_{n+1} = x\right) \geq 1 - \alpha - \alpha_{\text{i.i.d.}} - \text{CCG}(P, Q, x) \tag{57}$$

To evaluate the expected conditional coverage across the test distribution, we take the expectation over $x \sim Q_X$ and obtain

$$\mathbb{E}_{x\sim Q_X}[\Pr(Y_{n+1} \in \mathcal{C}_{\text{CQR}}(X_{n+1})|X_{n+1} = x)] \geq 1 - \alpha - \alpha_{\text{i.i.d.}} - \text{ICG}(P, Q), \tag{58}$$

where the $\text{ICG}(P, Q)$ is defined in Eq. (7) as the expectation of $\text{CCG}(P, Q, x)$ over Q_X.

Using our bound on $\text{ICG}(P, Q)$ in terms of the Wasserstein distance $W(P_{XY}, Q_{XY})$ in Eq. (13), we obtain a bound on the expected conditional coverage under distribution shift:

$$\begin{aligned}&\mathbb{E}_{x\sim Q_X}[\Pr(Y_{n+1} \in \mathcal{C}_{\text{CQR}}(X_{n+1})|X_{n+1} = x)] \\ &\geq 1 - \alpha - \alpha_{\text{i.i.d.}} - \sqrt{2\kappa L}\,(\eta \cdot W(P_{XY}, Q_{XY}) + 1/4)\,.\end{aligned} \tag{59}$$

Finally, the transformation by Augmented BNF leads to a more robust prediction set $\mathcal{C}^{\text{aug}}_{\text{BNF}}(X_{n+1})$. Let $\overline{Q}_{XY} := {f^{\text{aug}}_{\theta}}_{\#} Q_{XY}$. Assume that the conditions of Theorem 3.3 hold for $(P_{XY}, \overline{Q}_{XY})$, and denote the corresponding constant by $\bar{\eta}$. Applying Eq. (13) to this pair yields

$$\begin{aligned}&\mathbb{E}_{x\sim Q_X}[\Pr\big(Y_{n+1} \in \mathcal{C}^{\text{aug}}_{\text{BNF}}(X_{n+1})|X_{n+1} = x\big)] \\ &\geq 1 - \alpha - \alpha_{\text{i.i.d.}} - \sqrt{2\kappa L}\,\big(\bar{\eta} \cdot W(P_{XY}, \overline{Q}_{XY}) + 1/4\big)\,.\end{aligned} \tag{60}$$

We denote $\alpha_{\text{trans}} := \sqrt{2\kappa L}(\bar{\eta} \cdot W(P_{XY}, \overline{Q}_{XY}) + 1/4)$ to quantify the remaining deviation induced by imperfect alignment between calibration and test distributions. Figure 4 shows that the proposed transformation model effectively approximates the CQR under the i.i.d. condition.

N Prediction efficiency under source conditioning

N.1 Prediction inefficiency by Augmented BNF

Augmented BNF uses Eq. (24) to obtain prediction sets on the test distribution Q_{XY}. During training, this augmented component ε of the Y branch $f^{\text{aug}}_{\theta_Y}$ in Eq. (23) is sampled from a single Gaussian distribution $\mathcal{N}(0, 1)$, making it independent of the training sample sources. As a result, the model learns a shared transformation for all training distributions D^k_{XY} for $k = 1, ..., K$ to align with the calibration distribution P_{XY}.

At test time, the augmentation distribution provides no source information to the label transformation. Sharing this transformation across sources can lead to larger prediction sets when their conditional label distributions differ substantially.

The prediction inefficiency is reflected in Figure 18. The Augmented BNF produces noticeably larger prediction sets on Traffic, U.S. ILI, and Fair Med. This is due to the substantial variation in the conditional label distributions $D^k_{Y|x}$ within the three settings, such as significantly different supports, which leads to enlarged prediction sets.

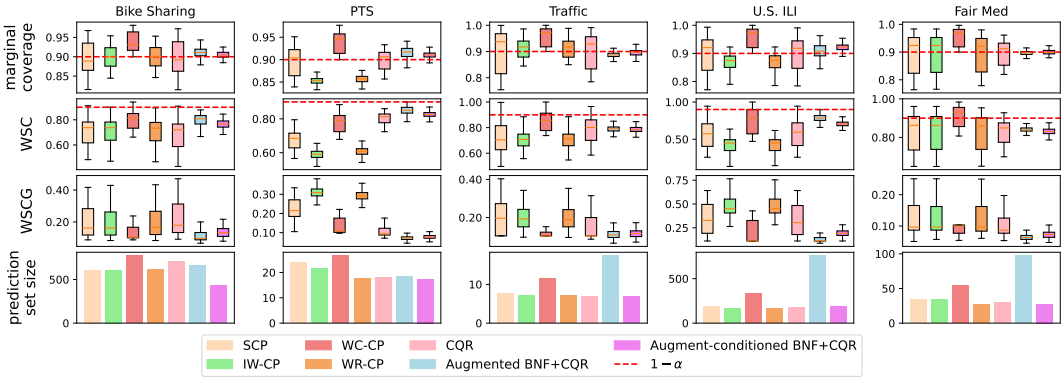


Figure 18: Prediction set size comparison. The standard Augmented BNF can produce large prediction sets, whereas the Augment-Conditioned variant significantly reduces set size while preserving coverage performance.

N.2 Efficient prediction by Augment-Conditioned BNF

To achieve smaller prediction sets, we propose **Augment-Conditioned BNF**, denoted as $f_\theta^{\text{aug-cond}}$. In this design, the augmented component ε is sampled from a distinct Gaussian distribution $\mathcal{N}^k$ if an instance (x, y) is from D_{XY}^k during training. In other words, ε serves not only to enhance expressiveness but also as a conditioning variable in $f_{\theta_Y}^{\text{aug-cond}}(y, \varepsilon)$. Consequently, during inference, we sample $\varepsilon_{n+1} \sim \mathcal{N}^k$ using the known source index k of the test sample, allowing us to construct a smaller prediction set $\mathcal{C}_{\text{BNF}}^{\text{aug-cond}}(X_{n+1})$. Using a similar calculation as in Eq. (24), we derive

$$\mathcal{C}_{\text{BNF}}^{\text{aug-cond}}(X_{n+1}) = \left\{ y : f_{\theta_Y}^{\text{aug-cond}}(y; \varepsilon_{n+1}) \in \mathcal{C}_{\text{C}}(\overline{X}_{n+1}) \right\}, \text{ where } \overline{X}_{n+1} = f_{\theta_X}(X_{n+1}) \tag{61}$$

Under exact joint alignment of each source with P_{XY} using its corresponding noise distribution $\mathcal{N}_k$, this set inherits the population conditional coverage of $\mathcal{C}_C$ in Eq. (3), after averaging over $\epsilon_{n+1} \sim \mathcal{N}_k$.

As shown in Figure 18, Augment-Conditioned BNF leads to significantly smaller prediction sets compared to the standard Augmented BNF. [53] also demonstrates that knowledge of the test sample's source can serve as valuable side information to improve prediction efficiency. Such side information is often available in real-world applications. For instance, in multi-center healthcare settings, models are required to generalize across different hospitals. In these scenarios, the center at which a patient is admitted is typically known and can be used to tailor the prediction procedure. Leveraging this information allows the model to generate tighter prediction sets while maintaining valid coverage guarantees.

We observe that the Augment-Conditioned BNF may exhibit slightly higher WSCG compared to the standard Augmented BNF, such as on U.S. ILI. That is attributed to data sparsity, caused by the use of more distinct Gaussian distributions to model the data sources. As the number of Gaussians increases, the samples become more dispersed, making it more challenging to learn the underlying population distributions effectively. Hence, although the Augment-Conditioned BNF yields smaller prediction sets, it compromises robustness in coverage, revealing an inherent trade-off.

O Additional experiment results

O.1 Ablation study across confidence levels

Rather than fixing $1-\alpha = 0.9$, we explore the performance of BNF across different confidence levels. After averaging over all datasets and trials, WSCG of Augmented BNF+CQR and five baselines with $1-\alpha$ from 0.1 to 0.9 are illustrated in Figure 19.

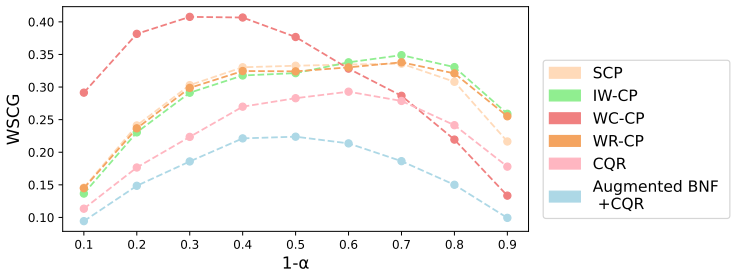


Figure 19: WSCG of Augmented BNF + CQR and baselines for $1-\alpha \in [0.1, 0.9]$, averaged across datasets.

It is shown that BNF consistently achieves the most robust conditional coverage with the lowest mean WSCG as the confidence level varies. Moreover, we can observe a pattern across all methods where WSCG tends to be higher in the middle confidence range and lower at both ends of the spectrum. At a high confidence level (i.e. $1-\alpha = 0.9$), prediction sets are so broad that they encompass the majority of possible outcomes, minimizing the potential for coverage gaps. Consequently, WSCG tends to be lower at high confidence levels. Similarly, prediction intervals are intentionally narrow at a low confidence level (i.e. $1-\alpha = 0.1$), capturing only a small subset of outcomes. This makes them inherently less sensitive to distribution shifts, thereby reducing coverage gaps. Hence, the coverage

gap is typically most pronounced at intermediate confidence levels, forming arch-like curves across the confidence spectrum in Figure 19.

O.2 Contribution of BNF and CQR

To examine the contributions of BNF and CQR, we compare SCP, CQR, Augmented BNF+SCP, and Augmented BNF+CQR using the same experimental protocol as the main results. Figure 20 summarizes marginal coverage, WSC, and WSCG at the nominal coverage level $1-\alpha = 0.9$, pooling test evaluations across the five experimental settings.

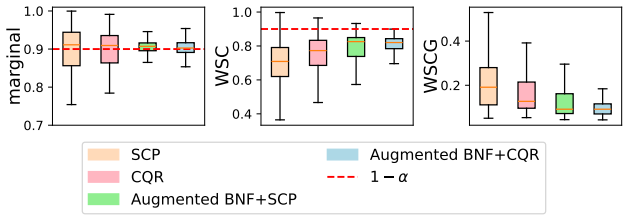


Figure 20: Ablation of BNF and CQR, with test evaluations pooled across the five experimental settings.

Adding Augmented BNF to SCP increases median WSC and reduces median WSCG, while concentrating marginal coverage near the nominal level. These improvements demonstrate that BNF is effective beyond its combination with CQR. Compared with this combination, Augmented BNF+CQR achieves similar median WSC and WSCG but narrower interquartile ranges for both metrics, suggesting more consistent conditional coverage performance across the pooled evaluations. These findings are consistent with the complementary roles of BNF and CQR: the former mitigates distribution shift, while the latter adapts prediction intervals to the input through conditional quantile estimation.

O.3 Challenges of feature conditioning in high dimensions

We present the experimental results of Feature-Conditioned BNF introduced in Section 7.1 across all datasets. This variant explicitly captures the dependency of Y on X. However, it increases the input dimension of $f^{\text{fea}}_{\theta_Y}$ to $d+1$, making the total input dimension of Feature-Conditioned BNF $2d+1$, where d is the dimension of X. As a result, the curse of dimensionality is exacerbated with a small sample-to-dimension ratio $|\mathcal{S}_{D^k}|/(2d+1)$, making true distributions harder to estimate. In contrast, Augmented BNF maintains a more favorable ratio of $|\mathcal{S}_{D^k}|/(d+2)$, as the input to $f^{\text{aug}}_{\theta_Y}$ is only two-dimensional. Consequently, with limited data, Feature-Conditioned BNF tends to yield higher WSCG due to poor approximation. We report the small sample-to-dimension ratio of Feature-Conditioned BNF in Table 4, and demonstrate its less robust conditional coverage in Figure 21, where its WSCG tends to be higher than that of Augmented BNF.

Meanwhile, although a one-dimensional projection may alleviate this issue, such dimensionality reduction inevitably discards information that can be crucial for accurately modeling the conditional calibration distribution, and its deterministic nature prevents it from achieving the level of expressiveness offered by Augmented BNF.

Table 4: Feature-Conditioned BNF holds a **small** sample-to-dimension ratio $|\mathcal{S}_{D^k}|/(2d+1)$.

Dataset	Bike Rental	PTS	Traffic	U.S. ILI	Fair Med
d	4	9	3	2	2
$\lvert\mathcal{S}_{D^k}\rvert$	2800	7500	2800	870	3000
$\lvert\mathcal{S}_{D^k}\rvert/(d+2)$	466.7	681.8	560.0	217.5	750.0
$\lvert\mathcal{S}_{D^k}\rvert/(2d+1)$	**311.1**	**394.7**	**400.0**	**174.0**	**600.0**

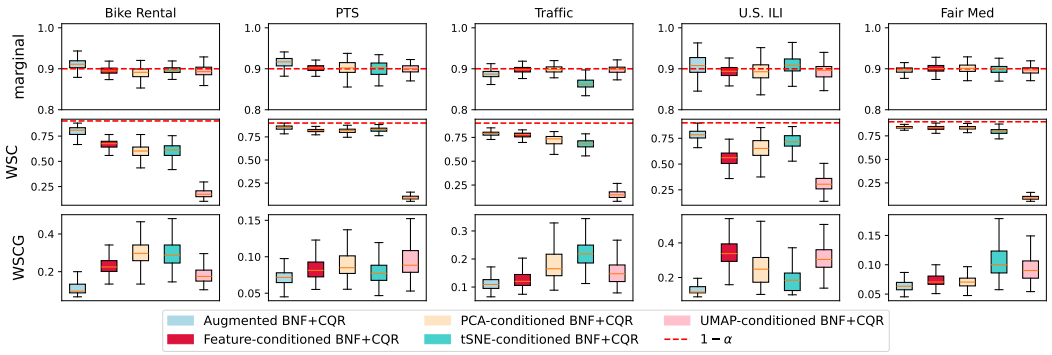


Figure 21: Comparison of conditioning the transformation of label Y on the feature and its one-dimensional X projection obtained via PCA, t-SNE, and UMAP.

O.4 Training Objectives and Test Mixtures

We examine how the training objective and test-mixture composition affect coverage and prediction efficiency. First, we compare average-source training with minimax training, which minimizes the maximum Wasserstein distance between the transformed source distributions and P_{XY}, using identical data splits and model settings. Second, keeping the average-source model fixed, we compare random-mixture testing with single-source testing, where each test set is drawn entirely from one randomly selected source.

Table 5: Comparison of training objectives and test settings for Augmented BNF+CQR at $1-\alpha = 0.9$. Results are aggregated across datasets and reported as mean $\pm$ standard deviation.

Objective	Test setting	Marginal	WSC	WSCG	Set size
Average	Random mixture	0.904 ± 0.018	0.811 ± 0.045	0.099 ± 0.038	311.0 ± 332.1
Minimax	Random mixture	0.886 ± 0.033	0.766 ± 0.093	0.141 ± 0.088	272.3 ± 302.6
Average	Single source	0.904 ± 0.021	0.812 ± 0.050	0.101 ± 0.041	330.8 ± 354.6

Table 5 shows that, under random-mixture testing, average-source training achieves marginal coverage closer to the target, higher WSC, and lower WSCG than minimax training, while minimax training produces smaller prediction sets. For the average-source model, single-source testing yields similar aggregate coverage metrics to random-mixture testing, with larger prediction sets on average.

The smaller prediction sets produced by minimax training in Table 5 come at the cost of undercoverage. This may reflect the instability of optimizing a hard maximum: gradients are propagated through the source with the largest current empirical Wasserstein distance, making updates sensitive to fluctuations in the source losses. In comparison, the average-source objective incorporates contributions from all sources, which may provide more stable optimization.

Single-source testing places all mixture mass on one source and therefore probes an extreme point of the source-mixture family. The similar aggregate coverage metrics suggest that the average-source objective remains empirically effective even in this concentrated setting. With the trained model held fixed, the larger average prediction-set size can reflect differences in the composition of the sampled test data, while aggregate coverage remains stable.

O.5 Distribution Shift via Label Perturbation

In Section 7.3, we evaluate the robustness of our method beyond the multi-source setting by introducing distribution shift based on label perturbation [42, 43]. Specifically, we consider a family of shifted distributions defined as

$$Q_{XY} \in \left\{ D_{XY} : (X,Y) \sim P_{XY},\ \delta \in [1, 1.5],\ \tilde{Y} = \delta Y,\ (X,\tilde{Y}) \sim D_{XY} \right\}.$$

This formulation corresponds to a multiplicative perturbation applied to the label, where the scaling factor δ is from $[1, 1.5]$. As a result, the marginal distribution of X remains unchanged, while the conditional distribution of $Y \mid X$ is perturbed.

Compared to additive perturbations, this scale-based transformation induces a more pronounced and heterogeneous shift, particularly for larger label values, thereby providing a challenging testbed for evaluating conditional coverage.

In our experiments, we construct five perturbed training environments for PTS and ten for each of the remaining settings, using label-scaling factors equally spaced over $[1, 1.5]$. BNF is trained to align these perturbed distributions with the original, unperturbed distribution P_{XY}. During inference, we evaluate the model on 100 shifted test distributions from the same perturbation family, independently sampling the scaling factor δ from the uniform distribution $\mathcal{U}(1, 1.5)$ for each test distribution.

The results across all datasets are presented in Figure 22, where our method demonstrates a favorable trade-off between robust conditional coverage and prediction efficiency.

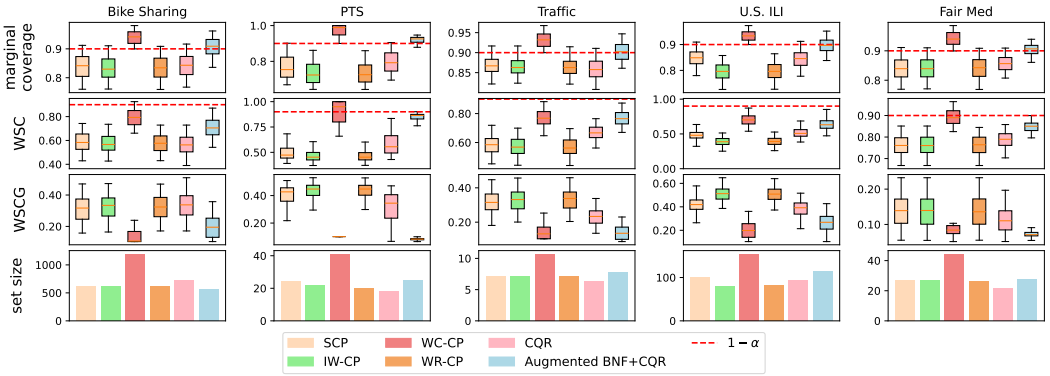


Figure 22: Distribution shift is induced by label perturbation. Compared with baselines, our method consistently maintains comparably robust marginal and conditional coverage without a high cost in prediction efficiency.

P Limitations

P.1 Stochastic prediction sets

One practical drawback of Augmented BNF lies in its reliance on stochastic augmentation through a random noise $\varepsilon \sim \mathcal{N}(0, 1)$, which is used to modulate the Y-branch of the Augmented BNF in Eq. (23). While this augmentation introduces flexibility, it also introduces randomness into the transformation. As a result, in Eq. (24), the prediction set produced by Augmented BNF for the same input x is no longer deterministic. The set varies across different forward passes depending on the realization of ε. This stochasticity undermines one of the appealing properties of standard conformal prediction: the deterministic and repeatable nature of the prediction set given a test point. In high-stakes domains, such randomness can lead to interpretability challenges or instability in downstream decisions.

A plausible solution is to perform multiple runs to approximate a stable prediction set, or to fix the random seed so that each test input yields a consistent prediction set across repeated queries.

P.2 Architectural incompatibility with one-dimensional features

Another limitation of Augmented BNF stems from its architectural dependency on Real NVP [28], a type of normalizing flow that is inherently designed for multi-dimensional transformations. Real NVP operates by alternating between dimensions of the input to apply affine coupling layers, as plotted in Figure 11. This necessitates a feature space $\mathcal{X}$ of at least two dimensions. Consequently, Augmented BNF inherits this constraint: its architecture presumes that the input feature x is multivariate.

In the case where x is one-dimensional, the affine coupling mechanism of Real NVP becomes undefined, rendering the model non-functional. As a result, Augmented BNF cannot be applied to

tasks with univariate inputs. This presents a clear barrier for applying to domains, where no natural multivariate feature exists. One might consider artificially expanding x with noise or engineered features to satisfy the dimensionality requirement, just like the augmented Y-branch in Eq. (23).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly reflect the paper's contribution.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We full discuss the limitations of the work in appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete proof of proposed theorems.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provided all necessary information to reproduce the result, including the architecture of the proposed branched structure.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We released the data and code publicly.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide all information, including datasets and models to understand the results.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We present coverage as boxplot and prediction set size as average over samples.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the computation cost of the proposed method with estimated time cost in appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: We do not anticipate direct societal impact from the proposed methodology or experiments, as the work is primarily methodological in nature.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite the original papers that produced the models and datasets we used in our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: the core method development in this research is about conformal prediction for regression tasks, and does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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