Accelerating Feature Conformal Prediction via Taylor Approximation

Zihao Tang[⋆] **Boyuan Wang**[◇] **Chuan Wen**[○] **Jiaye Teng**[⋆†]

[⋆]Shanghai University of Finance and Economics
[◇]Southern University of Science and Technology
[○]Shanghai Jiao Tong University

Abstract

Conformal prediction is widely adopted in uncertainty quantification, due to its post-hoc, distribution-free, and model-agnostic properties. In the realm of modern deep learning, researchers have proposed Feature Conformal Prediction (FCP), which deploys conformal prediction in a feature space, yielding reduced band lengths. However, the practical utility of FCP is limited due to the time-consuming non-linear operations required to transform confidence bands from feature space to output space. In this paper, we present Fast Feature Conformal Prediction (FFCP), a method that accelerates FCP by leveraging a first-order Taylor expansion to approximate these non-linear operations. The proposed FFCP introduces a novel non-conformity score that is both effective and efficient for real-world applications. Empirical validations showcase that FFCP performs comparably with FCP (both outperforming the Split CP version) while achieving a significant reduction in computational time by approximately 50x in both regression and classification tasks. The code is available at `https://github.com/ElvisWang1111/FastFeatureCP`.

1 Introduction

Machine learning has been successfully applied in numerous fields such as computer vision, natural language processing, and gaming [Jordan and Mitchell, 2015, Silver et al., 2017]. However, machine learning models usually suffer from overconfidence issues [Wei et al., 2022] and even hallucinations in large language models (LLMs) [Ji et al., 2023], which makes them unreliable and unable to be deployed in fields like finance and medicines [Gelijns et al., 2001, Thirumurthy et al., 2019, Morduch and Schneider, 2017]. Therefore, it is essential to develop techniques for uncertainty quantification and calibrate the original machine learning models Abdar et al. [2021], Guo et al. [2017], Chen et al. [2021], Gawlikowski et al. [2021].

Among the uncertainty quantification techniques, Conformal Prediction (Split CP, or split conformal prediction, Vovk et al. [2005]; Shafer and Vovk [2008b]; Burnaev and Vovk [2014]) stands out, because it is distribution-free, does not require retraining, and can be directly applied to various models. Conformal prediction deploys a calibration step to calibrate a base model and then construct the confidence band. The goal of conformal prediction is to return a band $\mathcal{C}_{1-\alpha}(X')$ such that

$$\mathbb{P}(Y' \in \mathcal{C}_{1-\alpha}(X')) \geq 1-\alpha, \tag{1}$$

where (X', Y') denotes a test point and $1-\alpha$ represents the confidence level.

In deep learning regimes, researchers try to utilize feature information in Split CP, since the feature space usually contains meaningful semantic information in neural networks [Shen et al., 2014]. This leads to Feature Conformal Prediction (FCP, Teng et al. [2022]).

[†]Correspondence to `tengjiaye@sufe.edu.cn`

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

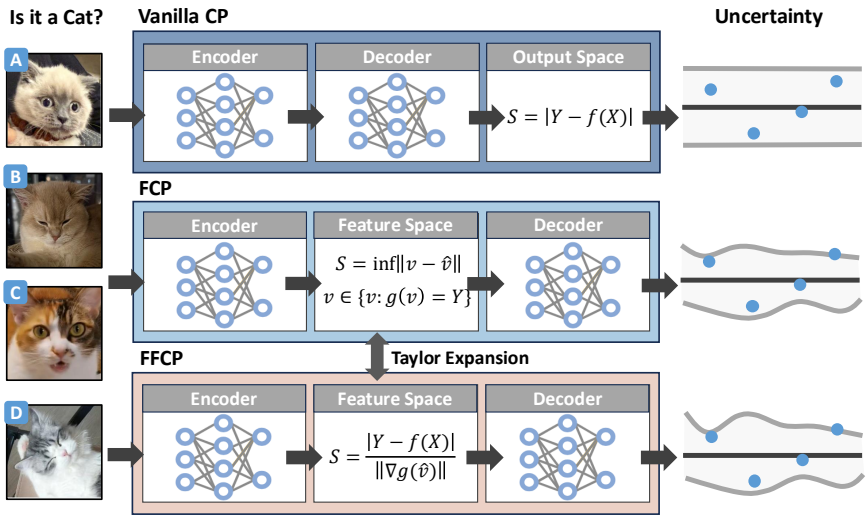


Figure 1: Comparison among Split CP, FCP, and FFCP. FCP and FFCP are more efficient compared to Split CP since they return different band lengths for different individuals. This is achieved by calculating a non-conformity score in the feature space. Besides, FFCP approximates FCP using a Taylor expansion, which leads to a different non-conformity score and accelerates the transformation from feature space to output space.

Fortunately, one may get different band lengths on different individuals by utilizing the feature information, leading to a shorter confidence band. As a comparison, Split CP only returns the same band length for all individuals in regression tasks, which indicates a longer length.

However, the practical applications of FCP are limited because (a) it is time-consuming, and (b) it only returns *estimated* bands on the output space, making it less efficient. These two issues come from the step *Band Estimation*, which transfers the confidence band from feature space to output space. This step involves complex non-linear operations called LiPRA [Xu et al., 2020] and therefore (a) the non-linear operation requires high computational costs, and (b) the configurations in LiPRA might finally influence the estimated band, further harming the performance of FCP.

In this paper, we present Fast Feature Conformal Prediction (FFCP), which offers a fast version for handling the aforementioned nonlinear operations in FCP. Different from Split CP and FCP, FFCP introduces a novel non-conformity score $s_{\text{ff}}(\cdot)$ that is simple to compute and does not require additional training,

$$s_{\text{ff}}(X, Y, g \circ h) = |Y - g \circ h(X)|/\|\nabla g(\hat{v})\|, \tag{2}$$

where (X, Y) denotes a sample, $g \circ h$ denotes a neural network with a feature layer h and a prediction head g, and the gradient $\nabla g(\hat{v})$ denotes the gradient of $g(\cdot)$ on the trained feature $\hat{v} \triangleq h(X)$, namely, $\nabla g(\hat{v}) = \frac{\mathrm{d}g \circ h(X)}{\mathrm{d}h(X)}$. We refer to Algorithm 2 for more details and illustrate the algorithm in Figure 1.

The above non-conformity score is closely related to FCP. Specifically, **FFCP with this non-conformity score can be regarded as a fast version of FCP, since it equivalently approximates the prediction head using a Taylor expansion**, which simplifies the aforementioned nonlinear operations. Fortunately, FFCP inherits the merits of FCP, for example, it also utilizes the semantic information in the feature.

From a theoretical perspective, we first demonstrate that FFCP is effective in Theorem 4.1, in that it returned a confidence band with empirical coverage larger than the given confidence $1 - \alpha$. Additionally, we demonstrate in Theorem 4.2 that FFCP produces a shorter confidence band than Split CP under a proposed square condition. The square conditions outline the properties of the feature space from two perspectives: expansion and quantile stability, implying that the feature space has a smaller distance between individual non-conformity scores and their quantiles. This reduces the cost of the quantile operation and therefore leads to a shorter confidence band. We also validate the square conditions using empirical observations.

From an empirical perspective, we conduct several experiments on real-world datasets and show that FFCP performs comparably with FCP, both outperforming Split CP, **while achieving nearly 50 times the speed of FCP in terms of runtime** for regression tasks. We further validate the approximation ability of FFCP with FCP using the correlation between the non-conformity score of FFCP and FCP. We also apply FFCP to the image segmentation problems to verify its general applications. Besides, we show that the concept in FFCP is pretty general, and can be combined with other variants of CP, *e.g.*, CQR [Romano et al., 2019a] and LCP [Guan, 2023] in regression tasks, and RAPS [Angelopoulos et al., 2020] in classification tasks.

Overall, our main contributions are summarized as follows:

- This work proposes FFCP, which serves as a fast version of FCP. FFCP achieves around 50x faster speed compared to FCP (Table 1) by utilizing Taylor expansions to approximate the prediction head in FCP. Besides, FFCP inherits the merits of FCP and efficiently exploits semantic information in the feature space.
- Theoretical insights demonstrate that FFCP returns shorter band length compared to Split CP (Theorem 4.2) while ensuring coverage exceeds the given confidence level under mild conditions (Theorem 4.1).
- Extensive experiments with both synthetic and real data demonstrate the effectiveness of the proposed FFCP algorithm (Table 2). Additionally, we demonstrate the universal applicability of our gradient-level techniques by extending them to other tasks such as classification (FFRAPS, Algorithm 5) and segmentation, and to various conformal prediction variants, including CQR (Algorithm 3) and LCP (Algorithm 4).

2 Related Work

Conformal prediction is a post-hoc calibration method dealing with uncertainty quantification [Vovk et al., 2005, Shafer and Vovk, 2008a, Barber et al., 2020], which is deployed in numerous fields [Ye et al., 2024, Kumar et al., 2023, Quach et al., 2023]. The variants of conformal prediction typically revolve around the concept of non-conformity scores, with four main branches of development.

Relaxing Exchangeability. The first branch relaxes the exchangeability requirement [Tibshirani et al., 2019, Hu and Lei, 2020, Podkopaev and Ramdas, 2021, Barber et al., 2022], leveraging weighted or reweighted quantiles to relax exchangeability. By doing so, it gains more flexibility and broader applicability in handling data that may not satisfy the standard exchangeability assumptions.

Diverse Structures. The second branch applies conformal prediction to various data structures, for example, classification tasks [Romano et al., 2020, Angelopoulos et al., 2020], time series data [Xu and Xie, 2021, Gibbs and Candès, 2021], censored data in survival analysis [Teng et al., 2021, Candès et al., 2023], high-dimensional data [Candès et al., 2021, Lei et al., 2013], Bellman-based data [Yang et al., 2024], counterfactuals and individual treatment effects [Lei and Candès, 2021], *etc*.

Another way involves model structures, such as k-NN regression [Papadopoulos et al., 2011a], quantiles incorporated [Romano et al., 2019b, Sesia and Candès, 2020], density estimators [Izbicki et al., 2020b], and conditional histograms [Sesia and Romano, 2021]. These methods further enrich the application scenarios of conformal prediction by adapting it to diverse model frameworks.

Enhancing Methods. The third branch focuses on enhancing the original conformal prediction with band length. Izbicki et al. [2020a] introduce CD-split and HPD-split methods, and Yang and Kuchibhotla [2021] develop selection methods to minimize band length. Of particular note is feature conformal prediction [Teng et al., 2022], which leverages neural network training information via feature spaces to improve band length.

Localized Conformal Prediction. The fourth branch focuses on enhancing the non-conformity score normalization by incorporating difficulty-related terms like $\frac{\|Y-\hat{Y}\|}{\sigma(X')}$ where Y denotes the true label, $\hat{Y}$ denotes the predicted label, and $\sigma(X')$ denotes the standard deviation related to X'. Here are three key approaches:

(1) Weight Adjustment via Calibration Distances. This approach calculates the distance from the test point to the calibration points and then uses these distances to define the weights of non-conformity scores in the calibration process [Han et al., 2022, Guan, 2023]. Our gradient-level techniques can be used to combine with this branch (see FFLCP in Algorithm 4).

(2) Normalization Using Proximity to Training Set. This approach utilizes the observation that a testing point exhibits smaller uncertainty when it is close to the training set, and uses such metrics to approximate $\sigma(X')$ [Papadopoulos et al., 2008, 2011b, Papadopoulos and Haralambous, 2011]. In the deep learning regimes, we believe that such procedures can be further improved by calculating the distance in the feature space rather than the input space, since feature layers usually contain more semantic information.

(3) Modeling $\sigma(X')$. This approach trains a separate model to estimate $\sigma(X')$, thereby enhancing the accuracy and adaptability of CP [Seedat et al., 2023, 2024]. However, this line of work heavily relies on the model performance and incurs high computational costs. As a comparison, our approach does not require additional training procedures.

Uncertainty Quantification. Uncertainty quantification is one of the most fundamental questions in machine learning. In addition to conformal prediction, many other approaches exist for quantifying uncertainty, including calibration-based techniques [Guo et al., 2017, Kuleshov et al., 2018, Nixon et al., 2019, Abdar et al., 2021, Chang et al., 2024] and Bayesian-based techniques [Blundell et al., 2015, Hernández-Lobato and Adams, 2015, Li and Gal, 2017, Izmailov et al., 2021, Jospin et al., 2022].

3 Preliminaries

We begin by introducing a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i\in[n]}$ indexed by $\mathcal{I}$. We split the dataset into two folds: a training fold $\mathcal{D}_{\text{tra}}$ indexed by $\mathcal{I}_{\text{tra}}$, and a calibration fold $\mathcal{D}_{\text{cal}}$ indexed by $\mathcal{I}_{\text{cal}}$. Denote the testing point by (X', Y'). For the model part, define f as a neural network. We partition $f = g \circ h$, where h denotes the feature function (the initial layers of the neural network) and g denotes the prediction head. For a sample (X, Y), we define $\hat{v} = h(X)$ as the trained feature. We follow the ideas in Teng et al. [2022] and define the surrogate feature as any feature v such that $g(v) = Y$.

Assumption 1 (exchangeability)**.** *Assume that the calibration data $(X_i, Y_i) \in \mathcal{D}_{cal}$ and the testing point (X', Y') are exchangeable. Formally, define $Z_i, i = 1, \ldots, |\mathcal{I}_{cal}| + 1$, as the above data pair. Then Z_i are exchangeable if arbitrary permutation follows the same distribution, i.e.,*

$$(Z_1, \ldots, Z_{|\mathcal{I}_{cal}|+1}) \overset{d}{=} (Z_{\pi(1)}, \ldots, Z_{\pi(|\mathcal{I}_{cal}|+1)}), \tag{3}$$

with arbitrary permutation π over $\{1, \cdots, |\mathcal{I}_{cal}| + 1\}$.

Typically, Split CP is composed of three key steps.
I. Training Step. We first train a base model using the training fold $\mathcal{D}_{\text{tra}}$.
II. Calibration Step. We calculate a non-conformity score $R_i = |Y_i - f(X_i)|$ using the calibration fold $\mathcal{D}_{\text{cal}}$. The form of the score function might vary case by case, quantifying the divergence between ground truth and predicted values.
III. Testing Step. We construct the confidence band for the testing point (X', Y') using the quantile of the non-conformity score $Q_{1-\alpha}$.

We present Split CP* in Algorithm 1, and provide its theoretical guarantee in Theorem 3.1.

Theorem 3.1. *Under Assumption 1, the confidence band $\mathcal{C}_{1-\alpha}(X')$ returned by Algorithm 1 satisfies*

$$\mathbb{P}(Y' \in \mathcal{C}_{1-\alpha}(X')) \geq 1 - \alpha. \tag{4}$$

4 Methodology

In this section, we first illustrate the motivation behind FFCP in Section 4.1. Specifically, we address the complexity of non-linear operators in FCP and derive FFCP from FCP. We then formally present the specific form of FFCP, including the non-conformity score, the returned bands, and the corresponding pseudocode. We finally provide theoretical analyses on the coverage and band length in Section 4.2.

*δ represents the Dirac function.

Algorithm 1 Split Conformal Prediction

Input: Confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\mathcal{I}}$, testing point X'

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tra}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{tra}}}$ and a calibration fold $\mathcal{D}_{\text{cal}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$;
2: Train a base model $f(\cdot)$ with training fold $\mathcal{D}_{\text{tra}}$;
3: For each $i \in \mathcal{I}_{\text{cal}}$, calculate the non-conformity score $R_i = |Y_i - f(X_i)|$;
4: Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{cal}}|+1}\sum_{i\in\mathcal{I}_{\text{cal}}}\delta_{R_i}+\delta_\infty$.

Output: $\mathcal{C}^{\text{Splitcp}}_{1-\alpha}(X') = [f(X') - Q_{1-\alpha}, f(X') + Q_{1-\alpha}]$.

Algorithm 2 Fast Feature Conformal Prediction

Input: Confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\mathcal{I}}$, testing point X'

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tra}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{tra}}}$ and a calibration fold $\mathcal{D}_{\text{cal}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$;
2: Train a base neural network with training fold $f(\cdot) = g \circ h(\cdot)$ with training fold $\mathcal{D}_{\text{tra}}$;
3: For each $i \in \mathcal{I}_{\text{cal}}$, calculate the non-conformity score $\tilde{R}_i = |Y_i - f(X_i)|/\|\nabla g(\hat{v}_i)\|$, where $\nabla g(\hat{v}_i)$ denotes the gradient of $g(\cdot)$ on the feature $\hat{v}_i \triangleq h(X_i)$, namely $\nabla g(\hat{v}_i) = \frac{\mathrm{d}g\circ h(X_i)}{\mathrm{d}h(X_i)}$;
4: Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{cal}}|+1}\sum_{i\in\mathcal{I}_{\text{cal}}}\delta_{\tilde{R}_i}+\delta_\infty$;

Output: $\mathcal{C}^{\text{ffcp}}_{1-\alpha}(X') = [f(X') - \|\nabla g(\hat{v}')\|Q_{1-\alpha},\ \ f(X') + \|\nabla g(\hat{v}')\|Q_{1-\alpha}]$, where $\hat{v}' = h(X')$.

4.1 Relationship between FFCP and FCP

In this section, we discuss the motivation behind FFCP. FFCP is inspired by FCP [Teng et al., 2022], which conducts conformal prediction in the feature space. However, since the band is constructed in the feature space, FCP requires a *Band Estimation* process to go from feature space to output space. Specifically, FCP applies *LiPRA* [Xu et al., 2020] which derives the band in the output space $\{g(v) : \|v - \hat{v}\| \le Q_{1-\alpha}\}$. Unfortunately, the exact band is difficult to represent explicitly since the prediction head g is usually highly non-linear, thereby resulting in significant computational complexity in terms of time. Therefore, we propose approximating g using a first-order Taylor expansion to simplify the aforementioned non-linear operator. The core steps of FCP include (a) calculating the non-conformity score (from output space to feature space), followed by (b) deriving the confidence band (from feature space to output space). We next introduce the concrete formulation of how FFCP approximates FCP.

From output space to feature space. FCP uses the non-conformity score $s_{\text{f}}(\cdot)$ in the feature space:

$$s_{\text{f}}(X, Y, g \circ h) = \inf_{v\in\{v:g(v)=Y\}} \|v - \hat{v}\|. \tag{5}$$

By using the Taylor expansion, one approximates g with $g(v) \approx g(\hat{v}) + \nabla g(\hat{v})(v - \hat{v})$. Plugging into the approximation of g leads to a new non-conformity score $s_{\text{ff}}(\cdot)$

$$s_{\text{ff}}(X, Y, g \circ h) = |Y - f(X)|/\|\nabla g(\hat{v})\|, \tag{6}$$

where $\nabla g(\hat{v})$ denotes the gradient of $g(\hat{v})$ on the feature $\hat{v}$, namely $\nabla g(\hat{v}) = \frac{\mathrm{d}g\circ h(X)}{\mathrm{d}h(X)}$.

From feature space to output space. After constructing the confidence band in the feature space, FCP maps this band to the output space. Specifically, FCP derives the following band in the output space which is called *Band Estimation*:

$$\{g(v) : \|v - \hat{v}\| \le Q_{1-\alpha}\}. \tag{7}$$

FCP proposes to use LiPRA in this process, which is time-consuming. By plugging into the Taylor approximation of g, one can construct the band $\mathcal{C}^{\text{ffcp}}_{1-\alpha}$ as

$$\mathcal{C}^{\text{ffcp}}_{1-\alpha}(X) = [g(\hat{v}) - \|\nabla g(\hat{v})\|Q_{1-\alpha}, g(\hat{v}) + \|\nabla g(\hat{v})\|Q_{1-\alpha}]\,. \tag{8}$$

Remark 1 (High-dimensional Response)**.** *When the response $Y_i = [Y_i^1, Y_i^2, \ldots, Y_i^m], i = 1, 2, \cdots, n$ is high-dimensional, one can deploy conformal prediction at a coordinate-wise level.*

Specifically, for dimension $j \in [m]$, we define the non-conformity score as

$$s^j_{\mathit{ff}}(X_i, Y_i, g \circ h) = |Y_i^j - f(X_i)^j| / \|\nabla g(\hat{v}_i)^j\|, \tag{9}$$

where $\nabla g(\hat{v}_i)^j = \left(\frac{\partial f(X_i)}{\partial h(X_i)}\right)_j$ represents the j-th row of the Jacobian matrix of f with respect to h at X_i.

We then compute a single quantile $Q_{1-\alpha}$ shared across all dimensions, defined by aggregating non-conformity scores from all coordinates and samples in the calibration set:

$$Q_{1-\alpha} = \mathit{Quantile}\left(\left\{s^j_{\mathit{ff}}(X_i, Y_i, g \circ h) : i \in \mathcal{I}_{\mathit{cal}}, j \in [m]\right\}, 1-\alpha\right). \tag{10}$$

The resulting confidence band for the j-th coordinate at test point X_i is given by:

$$\mathcal{C}^{\mathit{ffcp}}_{1-\alpha}(X_i)_j = \left[g(\hat{v}_i)^j - \|\nabla g(\hat{v}_i)^j\| Q_{1-\alpha},\ g(\hat{v}_i)^j + \|\nabla g(\hat{v}_i)^j\| Q_{1-\alpha}\right]. \tag{11}$$

Based on the above discussion, we present the full algorithm in Algorithm 2. Notably, the Taylor expansion in FFCP is usually different for each sample X, Y, which further leads to confidence bands that are individually different. Besides, FFCP inherits the advantages of FCP. For example, this framework is pretty general and can be combined with other variants of Split CP, *e.g.*, CQR [Romano et al., 2019a].

4.2 Theoretical Guarantee for FFCP

This section outlines the theoretical guarantee for FFCP concerning coverage (effectiveness) and band length (efficiency). Below, we offer the main theorems and defer the full proofs to Appendix A.1 and A.2. We first demonstrate that the confidence band produced by Algorithm 2 is valid under Assumption 1.

Theorem 4.1 (Coverage)**.** *Under Assumption 1, for any $\alpha > 0$, the confidence band returned by Algorithm 2 satisfies:*

$$\mathbb{P}(Y' \in \mathcal{C}^{\mathit{ffcp}}_{1-\alpha}(X')) \geq 1 - \alpha, \tag{12}$$

where the probability is taken over the calibration fold and the testing point (X', Y').

Next, we show that FFCP is provably more efficient than the Split CP. To simplify the discussion, we present an informal version of Theorem 4.2 here and postpone the formal version to Theorem A.1.

Theorem 4.2 (Band Length)**.** *Under mild assumptions, if the following square conditions hold:*

1. ***Expansion.*** *The feature space expands the differences between individual lengths and their quantiles.*
2. ***Quantile Stability.*** *Given a calibration set $\mathcal{D}_{\mathit{cal}}$, the quantile of the band length is stable in both feature space and output space.*

Then FFCP provably outperforms Split CP in terms of average band length.

By Theorem 4.2, FFCP is guaranteed to achieve a smaller average band length than Split CP. The *square conditions* imply that the feature space has a smaller distance between individual non-conformity scores and their quantiles. This reduction in the computational overhead of the quantile operation subsequently yields a shorter band length. We provide empirical verifications on this assumption, see Figure 4 for more details. The intuition behind Theorem 4.2 is as follows: Initially, FFCP and Split CP perform quantile operations in different spaces, with the *Expansion* condition ensuring that the quantile step in FFCP costs less. The ultimate *Quantile Stability* condition confirms that the band can be generalized from the calibration folds to the test folds.

5 Experiments

This section presents the experiments to validate the utility of FFCP. Firstly, we detail the experimental setup in Section 5.1. Secondly, we present that FFCP achieves both effectiveness and efficiency with

Table 1: Time comparison among Split CP, FCP, and FFCP. FFCP ensures faster running speed compared to FCP. The last column represents the speed improvement factor of FFCP compared to FCP. The time unit is in seconds.

DATASET	SPLIT CP	FCP	FFCP	FASTER
SYNTHETIC	$0.0088_{\pm 0.0003}$	$3.8939_{\pm 0.3725}$	$0.0902_{\pm 0.0056}$	43X
COM	$0.0047_{\pm 0.0010}$	$4.9804_{\pm 0.8588}$	$0.0844_{\pm 0.0187}$	59X
FB1	$0.0245_{\pm 0.0059}$	$5.9822_{\pm 0.9871}$	$0.1940_{\pm 0.0564}$	31X
FB2	$0.0414_{\pm 0.0070}$	$9.3534_{\pm 0.0927}$	$0.2510_{\pm 0.0058}$	37X
MEPS19	$0.0106_{\pm 0.0010}$	$3.3237_{\pm 0.0431}$	$0.0755_{\pm 0.0037}$	44X
MEPS20	$0.0152_{\pm 0.0016}$	$5.4003_{\pm 0.3945}$	$0.0948_{\pm 0.0077}$	57X
MEPS21	$0.0137_{\pm 0.0008}$	$4.1657_{\pm 0.0670}$	$0.0854_{\pm 0.0146}$	49X
STAR	$0.0030_{\pm 0.0006}$	$3.5842_{\pm 0.3722}$	$0.0332_{\pm 0.0066}$	108X
BIO	$0.0291_{\pm 0.0053}$	$7.5417_{\pm 1.1028}$	$0.2042_{\pm 0.0344}$	37X
BLOG	$0.0340_{\pm 0.0024}$	$8.0913_{\pm 1.2072}$	$0.2239_{\pm 0.0261}$	36X
BIKE	$0.0072_{\pm 0.0007}$	$3.5806_{\pm 0.0285}$	$0.0534_{\pm 0.0021}$	67X

Table 2: Comparisons of coverage and band length among Split CP, FCP, and FFCP. FFCP runs faster while performing comparably to FCP in most datasets and outperforming Split CP. For FFCP, we select the shortest band length among all layers.

METHOD	SPLIT CP		FCP		FFCP	
DATASET	COVERAGE	LENGTH	COVERAGE	LENGTH	COVERAGE	LENGTH
SYNTHETIC	$90.080_{\pm 0.951}$	$0.176_{\pm 0.015}$	$89.930_{\pm 0.956}$	$\mathbf{0.081}_{\pm 0.041}$	$90.080_{\pm 0.951}$	$0.176_{\pm 0.015}$
COM	$89.875_{\pm 0.985}$	$1.974_{\pm 0.071}$	$89.724_{\pm 1.087}$	$1.939_{\pm 1.408}$	$90.226_{\pm 2.179}$	$\mathbf{1.838}_{\pm 0.180}$
FB1	$90.254_{\pm 0.170}$	$2.004_{\pm 0.191}$	$90.198_{\pm 0.207}$	$2.010_{\pm 0.182}$	$90.168_{\pm 0.220}$	$\mathbf{1.472}_{\pm 0.232}$
FB2	$89.933_{\pm 0.206}$	$2.016_{\pm 0.218}$	$89.966_{\pm 0.130}$	$\mathbf{1.371}_{\pm 0.370}$	$89.868_{\pm 0.062}$	$1.425_{\pm 0.109}$
MEPS19	$90.567_{\pm 0.311}$	$3.982_{\pm 0.614}$	$90.605_{\pm 0.340}$	$3.493_{\pm 2.734}$	$90.352_{\pm 0.469}$	$\mathbf{3.134}_{\pm 0.309}$
MEPS20	$89.923_{\pm 0.715}$	$4.184_{\pm 0.316}$	$89.929_{\pm 0.770}$	$\mathbf{2.730}_{\pm 0.962}$	$89.615_{\pm 0.661}$	$3.268_{\pm 0.283}$
MEPS21	$90.019_{\pm 0.341}$	$3.732_{\pm 0.555}$	$90.038_{\pm 0.303}$	$3.393_{\pm 1.313}$	$89.745_{\pm 0.344}$	$\mathbf{3.146}_{\pm 0.506}$
STAR	$90.393_{\pm 1.494}$	$0.208_{\pm 0.004}$	$90.300_{\pm 1.362}$	$\mathbf{0.174}_{\pm 0.038}$	$90.393_{\pm 1.494}$	$0.208_{\pm 0.004}$
BIO	$89.875_{\pm 0.488}$	$1.661_{\pm 0.019}$	$89.930_{\pm 0.501}$	$\mathbf{1.412}_{\pm 0.265}$	$89.875_{\pm 0.488}$	$1.661_{\pm 0.019}$
BLOG	$90.176_{\pm 0.241}$	$3.524_{\pm 0.850}$	$90.151_{\pm 0.405}$	$2.795_{\pm 1.385}$	$90.059_{\pm 0.101}$	$\mathbf{2.741}_{\pm 0.517}$
BIKE	$89.871_{\pm 0.568}$	$0.703_{\pm 0.016}$	$89.394_{\pm 0.633}$	$2.147_{\pm 0.249}$	$89.624_{\pm 0.688}$	$\mathbf{0.635}_{\pm 0.030}$

faster execution in Section 5.2. Thirdly, in Section 5.3.1, we verify that FFCP can be easily deployed and performs robustly across various tasks, including classification and segmentation. Finally, in Section 5.3.2, we show that the gradient-level techniques used in FFCP can be extended to classic CP models such as CQR [Romano et al., 2019a] and LCP [Guan, 2023]. A more detailed account of this extension can be found in Section 5.3.

5.1 Experiments Setups

Datasets. We consider both synthetic datasets and realistic datasets, including **(a) synthetic dataset**: $Y = WX + \epsilon$, where $X \in [0,1]^{100}, Y \in \mathbb{R}, \epsilon \sim \mathcal{N}(0,1)$, W is a fixed random matrix. **(b) real-world unidimensional target datasets**: ten datasets from UCI machine learning [Asuncion, 2007] and other sources: community and crimes (*COM*), Facebook comment volume variants one and two (*FB1* and *FB2*), medical expenditure panel survey (*MEPS19–21*) [Cohen et al., 2009], Tennessee's student teacher achievement ratio (*STAR*) [Achilles et al., 2008], physicochemical properties of protein tertiary structure (*BIO*), blog feedback (*BLOG*) [Buza, 2014], and bike sharing (*BIKE*), **(c) real-world semantic segmentation dataset**: Cityscapes [Cordts et al., 2016], and **(d) real-world semantic classification dataset**: Imagenet-Val [Deng et al., 2009].

Algorithms. We compare three methods: Split CP, FCP, and FFCP, with Split CP serving as the baseline. For the one-dimensional scenario, we perform direct calculations. For higher-dimensional cases, we use a coordinate-wise level non-conformity score.

Evaluation. The algorithmic empirical performance is evaluated with the following metrics:

Table 3: Coverage and Band Length based on Gradient from Different Layers of Neural Networks. FFCP LAYER(·) represents using the gradient between the LAYER(·) and the input. The results in LAYER4 are equivalent to Split CP.

LAYER	LAYER1		LAYER2		LAYER3		LAYER4	
DATASET	COVERAGE	LENGTH	COVERAGE	LENGTH	COVERAGE	LENGTH	COVERAGE	LENGTH
SYNTHETIC	89.810±0.784	0.184±0.018	90.050±0.534	0.184±0.017	89.960±0.910	**0.182**±0.023	90.220±0.983	0.189±0.033
COM	90.476±1.889	1.878±0.224	90.226±2.179	**1.838**±0.180	89.674±1.465	1.853±0.136	89.825±0.646	2.037±0.188
FB1	90.112±0.199	3.540±0.327	90.212±0.357	2.860±0.327	90.083±0.216	1.597±0.052	90.168±0.220	**1.472**±0.232
FB2	89.953 ±0.250	3.530±0.384	89.897±0.235	3.048±0.510	89.956±0.159	2.077±0.517	89.868±0.062	**1.425**±0.109
MEPS19	90.155±0.643	3.251±0.396	90.352±0.469	**3.134**±0.309	90.440±0.183	3.184±0.482	90.586±0.246	3.795±0.640
MEPS20	89.934±0.520	4.302±1.377	89.889±0.621	3.573±0.488	89.615±0.661	**3.268**±0.283	89.82±0.689	3.817±0.308
MEPS21	89.496±0.262	3.443±0.487	89.623±0.275	3.218±0.239	89.745±0.344	**3.146**±0.506	90.026±0.301	3.452±0.711
STAR	90.901±1.732	0.221±0.002	90.993±1.807	0.217±0.003	91.039±1.442	0.210±0.004	90.300±1.248	**0.209**±0.004
BIO	89.937±0.391	2.292±0.077	90.022±0.375	2.042±0.067	89.991±0.594	2.080±0.063	90.127±0.476	**1.822**±0.025
BLOG	89.968±0.420	4.772±0.614	89.918±0.319	3.404±0.598	90.059±0.101	**2.741**±0.517	90.017±0.197	3.058±0.873
BIKE	89.917±0.791	1.701±0.254	89.568±0.476	1.138±0.114	89.495±0.579	0.794±0.068	89.624±0.688	**0.635**±0.030

- **Runtime** For runtime evaluation, the timing starts at the score calculation and ends with the final prediction bands returned. FFCP method records the total computation time for each layer, and then selects the layer that achieves the best results.
- **Coverage (Effectiveness)** Coverage refers to the observed frequency with which a test point falls within the predicted confidence interval. Ideally, a predictive inference method should yield a coverage rate slightly higher than $1-\alpha$ for a given significance level α.
- **Band length (Efficiency)** When the coverage exceeds $1-\alpha$, our goal is to minimize the length of the confidence band. For FFCP, since we use a 5-layers neural network, each layer can be viewed as a feature layer. Therefore, in the experiments, we obtain the band length returned by each of the 5 layers of the neural network. In the subsequent results, if only a single band length is presented, it corresponds to the shortest band length returned by the different neural network layers. Otherwise, the results for all layers from layer 0 to layer 4 (with the last layer typically representing the Split CP result) will be shown.

Let $Y = (Y^1, \ldots, Y^d) \in \mathbb{R}^d$ denote the d-dimensional response variable, and let $\mathcal{C}(X) \subseteq \mathbb{R}^d$ be the confidence band associated with predictor X. The length of this confidence band in each dimension is represented by the vector $\left(|\mathcal{C}(X)^1|, \ldots, |\mathcal{C}(X)^d|\right) \in \mathbb{R}^d$. Denote the indices of the test set by $\mathcal{I}_{\text{tes}}$ and the set of dimensions by $[d] = \{1, \ldots, d\}$. We then define the coverage and band length as:

$$\text{Coverage} = \frac{1}{|\mathcal{I}_{\text{tes}}|} \sum_{i \in \mathcal{I}_{\text{tes}}} \mathbb{I}\left(Y_i \in \mathcal{C}(X_i)\right), \quad \text{Band Length} = \frac{1}{|\mathcal{I}_{\text{tes}}|} \sum_{i \in \mathcal{I}_{\text{tes}}} \left(\frac{1}{d} \sum_{j=1}^{d} |\mathcal{C}(X_i)^j| \right), \quad (13)$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if its argument is true and 0 otherwise.

5.2 Results on Coverage, Band Length and Runtime

Runtime Comparison. The runtime comparison is presented in Table 1. The results show that FFCP outperforms FCP with an approximate 50x speedup in runtime. Notably, since Split CP is the most basic method and does not utilize additional tools, it exhibits the fastest runtime.

Coverage. Table 2 summarizes the coverage for the one-dimensional response. Experimental results indicate that the coverage of FFCP all exceeds the confidence level $1-\alpha$, affirming its effectiveness as stated in Theorem 4.1.

Band Length. The band length is detailed in Table 2 for a one-dimensional response. It is noteworthy that FFCP surpasses Split CP by achieving a shorter band length, thereby validating the efficiency of the algorithm.

5.3 Extensions of FFCP

This section provides the extensions of FFCP, which is divided into two parts. Section 5.3.1 mainly discusses the applications of FFCP beyond regression tasks, specifically in image classification [Angelopoulos et al., 2020] and segmentation tasks. Section 5.3.2 focuses on how the gradient-level

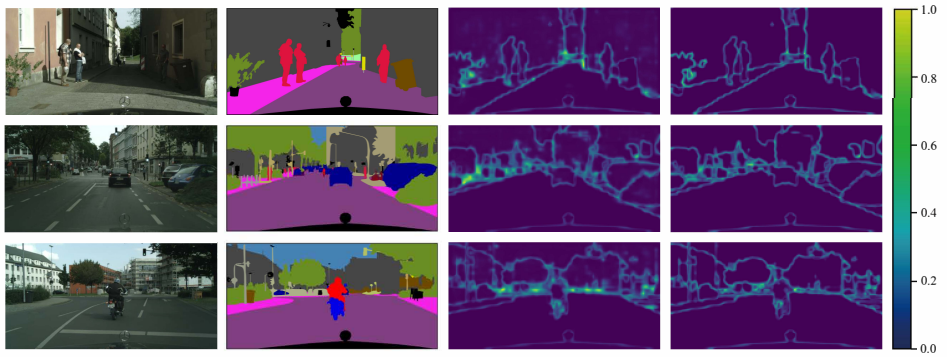


Figure 2: Segmentation uncertainty (FFCP vs. FCP). Both FFCP and FCP capture uncertainty concentrated around object boundaries, with FFCP producing more refined and sharper uncertainty bands. Notably, the two methods aim at slightly different goals: FCP operates at the image level, ensuring coverage for the entire predicted mask as a whole, while FFCP adopts a coordinate-wise approach, offering per-pixel statistical guarantees that better reflect local uncertainty.

techniques in FFCP can be extended to other CP variants, *e.g.*, CQR [Romano et al., 2019a] and LCP [Guan, 2023].

5.3.1 Other Tasks

Classification. We extend the FFCP techniques to classification tasks using the baseline RAPS [Angelopoulos et al., 2020] model, creating a new variant called FFRAPS (Fast Feature RAPS, Algorithm 5 in Appendix B.7). According to the experimental findings presented in Table 15, FFRAPS returns shorter band lengths while preserving the coverage compared to RAPS under most model structures.

Segmentation. The gradient-level techniques of FFCP also prove effective in segmentation tasks. The segmentation results in Figure 2 reveal that FFCP returns appropriate bands across different regions. Specifically, larger bands are observed in less informative areas, such as at object boundaries, whereas narrower bands are found in more informative regions. This validates the efficiency of FFCP in segmentation tasks.

5.3.2 Extending FFCP into Other Models

Conformalized Quantile Regression (CQR, Romano et al. [2019a]) The gradient-level techniques of FFCP are adaptable to other conformal prediction frameworks like CQR. We develop FFCQR (Fast Feature CQR, Algorithm 3 in Appendix B.5), which not only significantly reduces runtime compared to FCQR but also exhibits better performance than CQR. Additionally, we observe that for the neural network significant level setting $[\alpha, 1-\alpha]$ in the CQR method, as the α value increases, approaching $1-\alpha/2$, the performance of FFCQR gradually improves. For detailed experimental results in Table 8,9, 10, 11, 12 in the Appendix B.5.

Locally Adaptive Conformal Prediction (LCP, Guan [2023]) Integrating gradient-level techniques from FFCP into the LCP method leads to FFLCP (Fast Feature LCP, Algorithm 4 in Appendix B.6). Experimental results in Table 13 indicate that FFLCP outperforms LCP in terms of group coverage, highlighting an improvement in the adaptability of LCP to locally adaptive methods.

5.3.3 Comparison of FFCP with Other Baselines

Self-Supervised Conformal Prediction (SSCP, Seedat et al. [2023]) We additionally evaluate the recent feature-CP method SSCP in Appendix B.10. SSCP entails training two extra networks and is roughly 50× slower than FFCP. In all cases, FFCP also yields shorter bands, likely because SSCP depends more heavily on the base model's accuracy and on effective auxiliary-network training.

Full CP with CV+ [Barber et al., 2021] framework Although FFCP is a Split CP procedure, we also benchmark it against the full CP baseline CV+ and observe consistent advantages—most notably improved efficiency with comparable coverage. Complete experimental details and results are provided in the Appendix B.9.

6 Conclusion

In this paper, we propose FFCP, a gradient-based non-conformity score that is 50× faster than FCP. We establish its theoretical validity under mild assumptions and demonstrate its broad applicability across regression, classification, and segmentation tasks. We also introduce FFCQR and FFLCP, based on CQR and LCP, respectively. Finally, we evaluate FFCP in comparison with SSCP and the full CP baseline, CV+.

Although FFCP is gradient-based, it can be extended to settings where gradients are not available or are unreliable—*e.g.*, when gradients vanish or the function is non-differentiable—by incorporating zero-th order methods such as finite difference approximations or perturbation-based surrogates. This further broadens the applicability of our framework beyond differentiable models. For future work, the following points could be considered: (1) We use information from the first derivative and have not delved into higher-order derivatives, which may contain more feature information; (2) The gradient at a single point may be unstable, especially when the gradient is zero, so methods such as random smoothing could be considered.

References

M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

C. Achilles, H. P. Bain, F. Bellott, J. Boyd-Zaharias, J. Finn, J. Folger, J. Johnston, and E. Word. Tennessee's student teacher achievement ratio (star) project. *Harvard Dataverse*, 1:2008, 2008.

A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

A. U. Asuncion. Uci machine learning repository, university of california, irvine, school of information and computer sciences. 2007.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 2020.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1613–1622. JMLR.org, 2015.

E. Burnaev and V. Vovk. Efficiency of conformalized ridge regression. In *Conference on Learning Theory*, pages 605–622. PMLR, 2014.

K. Buza. Feedback prediction for blogs. In *Data analysis, machine learning and knowledge discovery*, pages 145–152. Springer, 2014.

E. Candès, L. Lei, and Z. Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

E. J. Candès, L. Lei, and Z. Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Y. Chen, D. Zhang, M. U. Gutmann, A. Courville, and Z. Zhu. Neural approximate sufficient statistics for implicit models. In *International Conference on Learning Representations*, 2021.

J. W. Cohen, S. B. Cohen, and J. S. Banthin. The medical expenditure panel survey: A national information resource to support healthcare cost research and inform policy and practice. *Medical Care*, 47:S44–S50, 2009.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. M. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu. A survey of uncertainty in deep neural networks. *CoRR*, abs/2107.03342, 2021.

A. C. Gelijns, J. G. Zivin, and R. R. Nelson. Uncertainty and technological change in medicine. *Journal of Health Politics, Policy and Law*, 26(5):913–924, 2001.

I. Gibbs and E. J. Candès. Adaptive conformal inference under distribution shift. In *NeurIPS*, 2021.

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

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017.

X. Han, Z. Tang, J. Ghosh, and Q. Liu. Split localized conformal prediction. *arXiv preprint arXiv:2206.13092*, 2022.

J. M. Hernández-Lobato and R. P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1861–1869. JMLR.org, 2015.

X. Hu and J. Lei. A distribution-free test of covariate shift using conformal prediction. *arXiv: Methodology*, 2020.

R. Izbicki, G. Shimizu, and R. B. Stern. Cd-split and hpd-split: efficient conformal regions in high dimensions. *arXiv preprint arXiv:2007.12778*, 2020a.

R. Izbicki, G. T. Shimizu, and R. B. Stern. Distribution-free conditional predictive bands using density estimators. *ArXiv*, abs/1910.05575, 2020b.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2): 29–48, 2022.

V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2801–2809. PMLR, 2018.

B. Kumar, C. Lu, G. Gupta, A. Palepu, D. Bellamy, R. Raskar, and A. Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

J. Lei, A. Rinaldo, and L. A. Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2013.

L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 2021.

Y. Li and Y. Gal. Dropout inference in bayesian neural networks with alpha-divergences. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2052–2061. PMLR, 2017.

J. Morduch and R. Schneider. *The financial diaries: How American families cope in a world of uncertainty*. Princeton University Press, 2017.

J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–41. Computer Vision Foundation / IEEE, 2019.

H. Papadopoulos and H. Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

H. Papadopoulos, A. Gammerman, and V. Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.

H. Papadopoulos, V. Vovk, and A. Gammerman. Regression conformal prediction with nearest neighbours. *J. Artif. Intell. Res.*, 40:815–840, 2011a.

H. Papadopoulos, V. Vovk, and A. Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011b.

A. Podkopaev and A. Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *UAI*, 2021.

V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019a.

Y. Romano, E. Patterson, and E. J. Candès. Conformalized quantile regression. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3538–3548, 2019b.

Y. Romano, M. Sesia, and E. J. Candès. Classification with valid and adaptive coverage. *arXiv: Methodology*, 2020.

N. Seedat, A. Jeffares, F. Imrie, and M. van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10160–10177. PMLR, 2023.

N. Seedat, J. Crabbé, Z. Qian, and M. van der Schaar. Triage: Characterizing and auditing training data for improved regression. *Advances in Neural Information Processing Systems*, 36, 2024.

M. Sesia and E. J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1): e261, 2020.

M. Sesia and Y. Romano. Conformal prediction using conditional histograms. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6304–6315, 2021.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, 2008a.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9 (3), 2008b.

Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*, pages 373–374, 2014.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359, 2017.

J. Teng, Z. Tan, and Y. Yuan. T-SCI: A two-stage conformal inference algorithm with guaranteed coverage for cox-mlp. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10203–10213. PMLR, 2021.

J. Teng, C. Wen, D. Zhang, Y. Bengio, Y. Gao, and Y. Yuan. Predictive inference with feature conformal prediction. *arXiv preprint arXiv:2210.00173*, 2022.

H. Thirumurthy, D. A. Asch, and K. G. Volpp. The uncertain effect of financial incentives to improve health behaviors. *Jama*, 321(15):1451–1452, 2019.

R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal prediction under covariate shift. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2526–2536, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

C. Xu and Y. Xie. Conformal prediction interval for dynamic time-series. In *ICML*, 2021.

K. Xu, Z. Shi, H. Zhang, Y. Wang, K.-W. Chang, M. Huang, B. Kailkhura, X. Lin, and C.-J. Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33:1129–1141, 2020.

Y. Yang and A. K. Kuchibhotla. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Z. Yang, E. Candès, and L. Lei. Bellman conformal inference: Calibrating prediction intervals for time series. *arXiv preprint arXiv:2402.05203*, 2024.

F. Ye, M. Yang, J. Pang, L. Wang, D. F. Wong, E. Yilmaz, S. Shi, and Z. Tu. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*, 2024.

Appendix

The complete proofs are presented in Section A, and the experiment details are outlined in Section B.

A Theoretical Proofs

We prove the theoretical guarantee for FFCP concerning coverage (effectiveness) in Section A.1 and band length (efficiency) in Section A.2.

A.1 Proofs of Theorem 4.1

The proof is based on the exchangeability of data (Assumption 1) on the calibration fold and test fold, hence the key step we need to derive is the exchangeability of the non-conformity scores $s_{\text{ff}}(X, Y, g \circ h) = |Y - f(X)|/\|\nabla g(\hat{v})\|$. We define the relevant symbols: $\mathcal{D}_{\text{tra}}$ represents the train fold, $\mathcal{D}_{\text{tes}}$ represents the test fold, $\mathcal{D}_{\text{cal}}$ represents the calibration fold, and $\mathcal{D}' = \{(X_i, Y_i)\}_{i\in[m]}$ is the intersection of the two folds. m is the number of data points in $\mathcal{D}'$.

Similar to Teng et al. [2022], we first prove that for any function $\tilde{h} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is independent of $\mathcal{D}'$, $\tilde{h}(X_i, Y_i)$ satisfies exchangeability. For the CDF F_R of $\tilde{h}$ and its perturbation CDF F_R^{π}, π is a random perturbation. We can conclude,

$$
\begin{aligned}
&F_R(u_1, \ldots, u_n \mid \mathcal{D}_{\text{tra}}) \\
=&\mathbb{P}(\tilde{h}(X_1, Y_1) \le u_1, \ldots, \tilde{h}(X_n, Y_n) \le u_n \mid \mathcal{D}_{\text{tra}}), \\
=&\mathbb{P}((X_1, Y_1) \in \mathcal{C}_{\tilde{h}^{-1}}(u_1-), \ldots, (X_n, Y_n) \in \mathcal{C}_{\tilde{h}^{-1}}(u_n-) \mid \mathcal{D}_{\text{tra}}), \\
=&\mathbb{P}((X_{\pi(1)}, Y_{\pi(1)}) \in \mathcal{C}_{\tilde{h}^{-1}}(u_1-), \ldots, (X_{\pi(n)}, Y_{\pi(n)}) \in \mathcal{C}_{\tilde{h}^{-1}}(u_n-) \mid \mathcal{D}_{\text{tra}}), \\
=&\mathbb{P}(\tilde{h}(X_{\pi(1)}, Y_{\pi(1)}) \le u_1, \ldots, \tilde{h}(X_{\pi(n)}, Y_{\pi(n)}) \le u_n \mid \mathcal{D}_{\text{tra}}), \\
=&F_R^{\pi}(u_1, \ldots, u_n \mid \mathcal{D}_{\text{tra}}),
\end{aligned}
\tag{14}
$$

where $\mathcal{C}_{\tilde{h}^{-1}}(u-) = \{(X, Y) : \tilde{h}(X, Y) \le u\}$.

Next, we need to show the non-conformity score function

$$
s_{\text{ff}}(X, Y, g \circ h) = |Y - f(X)|/\|\nabla g(\hat{v})\|, \tag{15}
$$

which is independent of the dataset $\mathcal{D}'$.

We can see that the non-conformity score $s_{\text{ff}}(X, Y, g \circ h)$ on $\mathcal{D}'$ uses information from g and h, both of which depend only on the training set $\mathcal{D}_{\text{tra}}$. Moreover, calculating this non-conformity score in the Algorithm 2 uses only single-point information, not the entire dataset $\mathcal{D}'$.

By integrating the aforementioned, we deduce that the non-conformity scores $s_{\text{ff}}(X, Y, g \circ h)$ on $\mathcal{D}'$ exhibit exchangeability. This exchangeability, as per Lemma 1 in Tibshirani et al. [2019], lends theoretical support to the efficacy of FFCP.

A.2 Proofs of Theorem 4.2

Our main conclusions are inspired by Theorem 4 in Teng et al. [2022]. The details are as follows

Definitions. Let $\mathcal{P}$ denote the overall population distribution. The calibration set $\mathcal{D}_{\text{cal}}$ consists of n samples drawn from $\mathcal{P}$. We denote the specific distribution of these samples as $\mathcal{P}^n$. The model under consideration, $f = g \circ h$, includes h as the feature extractor and g as the prediction head, with g assumed to be a continuous function. $V_{\mathcal{D}}^o$ represent the individual length in output space, given data set $\mathcal{D}$. The term $Q_{1-\alpha}(R)$ represents the $(1-\alpha)$-quantile of the set R, which is adjusted to include the value 0. Furthermore, $\mathbb{M}[\cdot]$ signifies the mean value of a set, and subtracting a real number from a set indicates that the subtraction is applied uniformly to all elements within the set.

Split CP. Let $V_{\mathcal{D}_{\text{cal}}}^o = \{v_i^o\}_{i\in\mathcal{I}_{\text{cal}}}$ denote the individual length in the output space for Split CP, given the calibration set $\mathcal{D}_{\text{cal}}$. Since Split CP returns band length with $1-\alpha$ quantile of non-conformity score, the resulting average band length is derived by $2Q_{1-\alpha}(V_{\mathcal{D}_{\text{cal}}}^o)$.

Fast Feature CP. According to the definition of FFCP, $V_{\mathcal{D}}^f = V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|$,

The resulting band length in FFCP is denoted by $2\mathbb{E}_{(X',Y')\sim\mathcal{P}}(\|\nabla g(\hat{v'})\| \cdot Q_{1-\alpha}(V_{\mathcal{D}_{\text{cal}}}^o/\|\nabla g(\hat{v}_{\text{cal}})\|)$.

Theorem A.1. *(FFCP is provably more efficient). Assume that the non-conformity score is in norm-type. We assume a Holder assumption that there exist $\alpha > 0, L > 0$ such that $|\mathcal{H}(x;X) - \mathcal{H}(y;X)| \leq L|x-y|^\alpha$ for all X, where $\mathcal{H}$ denotes the length of the prediction interval in the output space for samples. Then if the feature space satisfies the following square conditions:*

1. ***Expansion.*** *The feature space expands the differences between individual length and their quantiles, namely, $L\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\mathbb{M}|Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|) - V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\||^\alpha < \mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\mathbb{M}[Q_{1-\alpha}(V_{\mathcal{D}}^o) - V_{\mathcal{D}}^o] - 2\max\{L,1\}(c/\sqrt{n})^{\min\{\alpha,1\}}$.*
2. ***Quantile Stability.*** *Given a calibration set $\mathcal{D}_{cal}$, the quantile of the band length is stable in both feature space and output space, namely, $\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}|Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|) - Q_{1-\alpha}(V_{\mathcal{D}_{cal}}^o/\|g(\nabla\hat{v}_{cal})\|)| \leq \frac{c}{\sqrt{n}}$ and $\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}|Q_{1-\alpha}(V_{\mathcal{D}}^o) - Q_{1-\alpha}(V_{\mathcal{D}_{cal}}^o)| \leq \frac{c}{\sqrt{n}}$.*

Then FFCP provably outperforms Split CP in terms of average band length, namely,

$$\mathbb{E}_{(X',Y')\sim\mathcal{P}}(\|\nabla g(\hat{v'})\| \cdot Q_{1-\alpha}(V_{\mathcal{D}_{cal}}^o/\|\nabla g(\hat{v}_{cal})\|) < Q_{1-\alpha}(V_{D_{cal}}^0),$$

where the expectation is taken over the calibration fold and the testing point (X', Y').

Proof of Theorem A.1. We first proof with *Expansion* Assumption,

$$\begin{aligned} &L\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\mathbb{M}|Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|) - V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\||^\alpha < \mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\mathbb{M}[Q_{1-\alpha}(V_{\mathcal{D}}^o) - V_{\mathcal{D}}^o] \\ &- 2\max\{L,1\}(c/\sqrt{n})^{\min\{\alpha,1\}}. \end{aligned} \tag{16}$$

And we can obtain

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\mathbb{M}V_{\mathcal{D}}^o <& \mathbb{E}_{\mathcal{D}}Q_{1-\alpha}(V_{\mathcal{D}}^o) \\ &- 2\max\{L,1\}(c/\sqrt{n})^{\min\{\alpha,1\}} - L\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\mathbb{M}|Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|) - V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\||^\alpha. \end{aligned} \tag{17}$$

According to Holder condition for quantile function, we obtain that $\mathbb{M}(\|\nabla g(\hat{v})\| \cdot Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|))$
$\leq \mathbb{M}V_{\mathcal{D}}^o + L\mathbb{M}|Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|) - V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\||^\alpha$, therefore

$$\mathbb{E}_{\mathcal{D}}\mathbb{M}(\|\nabla g(\hat{v})\| \cdot Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|)) < \mathbb{E}_{\mathcal{D}}Q_{1-\alpha}(V_{\mathcal{D}}^o) - 2\max\{1,L\}[c/\sqrt{n}]^{\min\{1,\alpha\}}. \tag{18}$$

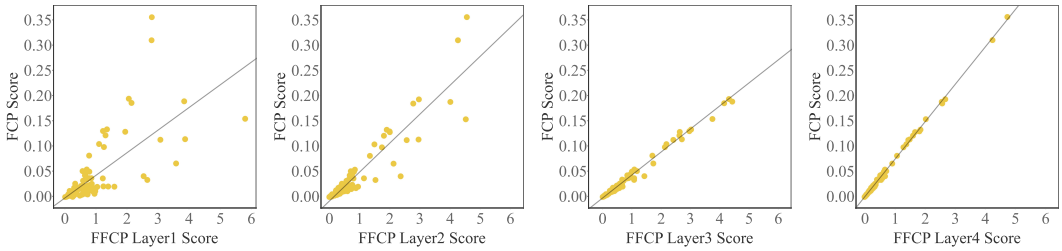
As the *Quantile Stability* assumption, we have that $\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}|Q_{1-\alpha}(V_{\mathcal{D}}^o/\|\nabla g(\hat{v})\|) - Q_{1-\alpha}(V_{\mathcal{D}_{\text{cal}}}^o/\|\nabla g(\hat{v}_{\text{cal}})\|)|$
$\leq \frac{c}{\sqrt{n}}$ and $\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}|Q_{1-\alpha}(V_{\mathcal{D}}^o) - Q_{1-\alpha}(V_{\mathcal{D}_{\text{cal}}}^o)| \leq \frac{c}{\sqrt{n}}$. Therefore,

$$\begin{aligned} &2\mathbb{E}(\|\nabla g(\hat{v})\| \cdot Q_{1-\alpha}(V_{\mathcal{D}_{\text{cal}}}^o/\|\nabla g(\hat{v}_{\text{cal}})\|) \\ <&2Q_{1-\alpha}(V_{\mathcal{D}}^o) - 2\max\{1,L\}[c/\sqrt{n}]^{\min\{1,\alpha\}}, \\ <&2Q_{1-\alpha}(V_{\mathcal{D}}^o). \end{aligned} \tag{19}$$

□

B Experimental Details

Section B.1 introduces the omitted experimental details. Section B.2 certifies the square conditions. Section B.3 discusses discusses the robustness of FFCP coverage with respect to the splitting point and across each network layer. Section B.4 demonstrates that FFCP performs similarly to Split CP in untrained neural networks, confirming that FFCP's efficiency is due to the semantic information trained in the feature space. Section B.5 proposes FFCQR after applying the gradient-level techniques of FFCP to CQR. Section B.6 proposes FFLCP after applying the gradient-level techniques of FFCP to LCP. Section B.7 proposes FFRAPS after applying the gradient-level techniques of FFCP to RAPS. Finally, Section B.8 provides additional experimental results.



(a) Layer1 (b) Layer2 (c) Layer3 (d) Layer4

Figure 3: Scatter plot of FCP Score and FFCP Score at different layers. The relationship between the FCP Score and the FFCP Score is positively correlated, which indicates that the FFCP Score effectively approximates the FCP Score.

B.1 Experimental Details

All the tests are performed on a desktop with an Intel Core i9-12900H CPU, NVIDIA GeForce RTX 4090 GPU, and 32 GB memory.

We conduct several more experiments in to establish the close relationship between FFCP and FCP, to demonstrate the benefits of FFCP from the good representation of gradient, and to provide empirical validations for the theoretical insights.

Model Architecture. For the one-dimensional we employ a four-layer neural network, with each layer consisting of 64 dimensions. For the semantic segmentation experiment, we utilize a network architecture combining ResNet50 with two additional convolutional layers. We use ResNet50 as the base feature extractor h, and the two subsequent convolution layers form the prediction head g.

Correlation Between FCP and FFCP Scores Across Layers. We compare the relationship between the scores of FCP and FFCP through experiments. Figure 3 indicates a positive correlation between the non-conformity scores of the two algorithms, suggesting that FFCP shares similarities with FCP in score function. This suggests that FFCP, while computationally efficient, provides a close approximation to FCP. The observed discrepancies may be attributed to the complex, layer-dependent non-linear transformations introduced by the decoder, under which the accuracy of the Taylor-based linear approximation tends to decline as the degree of non-linearity increases.

Decoder Runtime Comparison. We conducted additional experiments to evaluate how the decoder's layer depth influences computational efficiency. The running times presented below consider the decoder at various depths, explicitly showing the proportional relationship between decoder depth and computational cost.

In our primary comparison (FFCP vs. FCP), we maintained consistency by selecting the layer index as 2 for the FFCP, aligning it with the depth used by FCP. The detailed experimental results are summarized in Table 4:

Table 4: Running time comparison (mean $\pm$ std) for different decoder depths in STAR dataset.

Layer Depth	Split CP	FCP	FFCP	FASTER
2 Layer	$0.0044_{\pm 0.0009}$	$3.1302_{\pm 0.4795}$	$0.0247_{\pm 0.0053}$	127x
4 Layer	$0.0047_{\pm 0.0001}$	$5.8031_{\pm 0.2011}$	$0.0239_{\pm 0.0010}$	243x
6 Layer	$0.0058_{\pm 0.0003}$	$9.4411_{\pm 0.5270}$	$0.0277_{\pm 0.0011}$	341x
8 Layer	$0.0064_{\pm 0.0002}$	$12.2794_{\pm 0.1668}$	$0.0294_{\pm 0.0012}$	418x
10 Layer	$0.0072_{\pm 0.0001}$	$15.8716_{\pm 0.3363}$	$0.0335_{\pm 0.0015}$	474x
12 Layer	$0.0077_{\pm 0.0003}$	$19.4352_{\pm 0.8947}$	$0.0373_{\pm 0.0021}$	521x

These results demonstrate that while the computational cost of FCP grows significantly with the increasing depth of the decoder network, FFCP remains computationally efficient, only exhibiting marginal increases in runtime. Thus, FFCP provides substantial efficiency benefits, especially when employing deep network architectures.

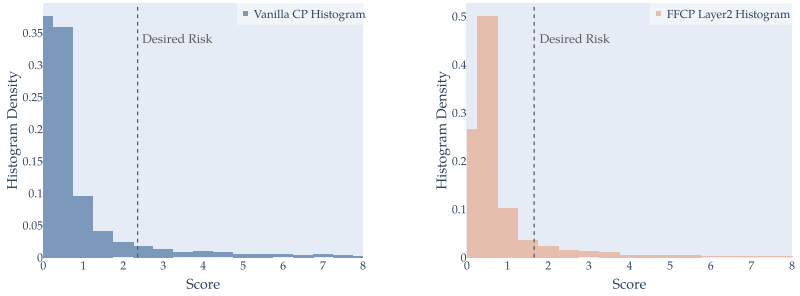
Robustness for FFCP. The empirical performance of FFCP demonstrates its robustness, as seen in the ablation studies on splitting points. We demonstrate that coverage remains robust across different splitting points in neural networks, as detailed in Table 5 in Appendix B.3. Furthermore, the results from different layers of the FFCP network are consistent, as presented in Table 3

B.2 Verifying Square Conditions

We verify the square conditions in this section. The key component of the square conditions is *Expansion* condition, which states that performing the quantile step does not result in a significant loss of efficiency.

For computational simplicity, We take exponent $\alpha = 1$ and do not consider the Lipschitz factor L. We next provide experiment results in Figure 4 on comparing the distribution of the scores between Split CP with FFCP.

From the figure, we observe that the overall distribution of FFCP non-conformity scores is closer to the quantile. This numerically validates that $\mathbf{M}\left|Q_{1-\alpha}(V^o_{\mathcal{D}}/|\nabla g(\hat{v})|) - V^o_{\mathcal{D}}/|\nabla g(\hat{v})|\right|$ is less than $\mathbf{M}\left[Q_{1-\alpha}(V^o_{\mathcal{D}}) - V^o_{\mathcal{D}}\right]$.



(a) Split CP (b) FFCP

Figure 4: Empirical validation of Theorem A.1. We plot the score distributions and their corresponding quantiles ($\alpha = 0.1$) of Split CP (left) and FFCP (right). Compared to Split CP, the non-conformity scores of FFCP are closer to their quantiles, leading to a shorter band. Compared to Split, FFCP exhibits a more stable distribution with higher quantiles, leading to better performance for FFCP. FFCP selects layer 2 for display.

B.3 Robustness of FFCP

To verify that the coverage by FFCP maintains its robustness despite changes in the splitting point, we performed a network split. The experimental results, detailed in Table 5, demonstrate that FFCP is indeed robust.

Table 5: Ablation study of the number of layers in h and g in unidimensional tasks. For the sake of avoiding redundancy, we set $\alpha = 0.05$.

DATASET		FACEBOOK1		MEPS19		BLOG	
METHOD	NUMBER($g \circ h$)	COVERAGE	LENGTH	COVERAGE	LENGTH	COVERAGE	LENGTH
SPLIT CP	/	95.24 ± 0.16	4.60 ± 0.50	95.35 ± 0.23	7.34 ± 1.01	95.08 ± 0.11	7.88 ± 0.97
FFCP	$h:0 \quad g:4$	94.88 ± 0.19	5.35 ± 0.42	95.18 ± 0.40	5.36 ± 0.52	94.97 ± 0.31	8.28 ± 0.59
	$h:1 \quad g:3$	94.84 ± 0.14	4.21 ± 0.46	95.13 ± 0.38	5.15 ± 0.49	94.99 ± 0.18	6.37 ± 1.22
	$h:2 \quad g:2$	95.14 ± 0.13	2.48 ± 0.09	95.19 ± 0.36	5.57 ± 0.87	95.08 ± 0.12	5.36 ± 0.93
	$h:3 \quad g:1$	95.16 ± 0.18	2.59 ± 0.72	95.34 ± 0.22	6.95 ± 1.13	95.05 ± 0.11	6.46 ± 1.52
	$h:4 \quad g:0$	95.24 ± 0.16	4.60 ± 0.50	95.35 ± 0.23	7.34 ± 1.01	95.08 ± 0.11	7.88 ± 0.97

Table 6: Untrained model comparison between Split CP and FFCP. When the model has not been sufficiently trained, FFCP performs similarly to Split CP. This means that the model's performance determines the quality of the feature information in the gradient. When the model performs poorly, the gradient information obtained by FFCP is inaccurate. On the other hand, this also suggests that FFCP effectively utilizes the feature information in the gradient when the model is well-trained.

METHOD	SPLIT CP		FFCP	
DATASET	COVERAGE	LENGTH	COVERAGE	LENGTH
SYNTHETIC	90.23±0.45	**2.34**±0.01	90.22±0.96	2.41±0.01
COM	90.33±1.81	4.86±0.13	90.43±1.99	**4.73**±0.08
FB1	90.18±0.19	3.57±0.09	90.10±0.13	**3.57**±0.08
FB2	90.16±0.11	3.66±0.11	90.12±0.14	**3.66**±0.06
MEPS19	90.80±0.43	**4.33**±0.07	90.85±0.58	4.38±0.07
MEPS20	90.15±0.55	**4.41**±0.23	90.27±0.63	4.46±0.25
MEPS21	89.80±0.45	**4.41**±0.17	89.89±0.56	**4.41**±0.15
STAR	89.79±0.51	**1.88**±0.01	89.98±0.56	1.94±0.01
BIO	90.16±0.20	4.09±0.02	90.07±0.14	**4.04**±0.02
BLOG	90.11±0.30	**2.53**±0.12	90.12±0.28	2.55±0.14
BIKE	89.55±0.82	**4.56**±0.09	89.57±0.86	4.60±0.10

Table 7: Untrained model comparison between Split CP and FFCP by Layer on Each Dataset (lower is better).

DATASET	LAYER1	LAYER2	LAYER3	LAYER4	LAYER5 (SPLIT)
SYNTHETIC	2.87±0.03	2.81±0.03	2.89±0.01	2.41±**0.01**	**2.34**±0.01
COM	5.41±0.14	5.30±0.18	5.03±0.09	**4.73**±**0.08**	4.86±0.13
FB1	3.90±0.01	3.77±0.10	3.62±0.08	3.56±0.08	**3.56**±**0.09**
FB2	4.02±0.09	3.90±0.8	3.74±0.08	3.66±0.06	**3.66**±**0.11**
MEPS19	4.38±0.06	4.41±0.07	4.42±0.07	4.38±0.07	**4.33**±**0.07**
MEPS20	4.43±0.21	4.44±0.23	4.48±0.25	4.46±0.25	4.41±0.23
MEPS21	**4.41**±**0.18**	4.46±0.18	4.48±0.18	4.41±0.15	4.41±0.17
STAR	2.49±0.06	2.40±0.05	2.15±0.03	1.94±0.01	**1.88**±**0.01**
BIO	4.27±0.02	4.14±0.02	4.07±0.02	**4.04**±**0.02**	4.08±0.02
BLOG	2.56±0.15	2.54±0.15	2.57±0.13	2.55±0.14	**2.53**±**0.12**
BIKE	4.69±0.09	4.67±0.07	4.57±0.09	4.60±0.10	**4.56**±**0.09**

B.4 FFCP works due to semantic information in feature space

One of our primary advantages is that FFCP leverages the semantic information of gradient in feature space. This is due to the fact that gradient-level techniques in feature space improve efficiency via the robust feature embedding abilities of well-trained neural networks.

On the other hand, when the base model is untrained and initialized randomly, lacking meaningful semantic representation in gradient, the band length produced by FFCP is comparable to Split CP. For results, see Table 6.

FFCP on untrained network. We propose that FFCP returns shorter band lengths through its deployment of deep representations from the gradients. To test this view, we contrast FFCP's performance using an untrained neural network against a baseline model. Using an incompletely trained neural network, FFCP's performance deteriorates and becomes comparable to that of Split CP. This is due to the partially incorrect semantic information in the gradient, which *misleads* FFCP. We defer the results to Table 6 and have updated the results in Table 7 by selecting the best-performing layer, which confirms that FFCP underperforms when the model is not sufficiently trained.

B.5 FFCQR

This section highlights the adaptability of FFCP's gradient-level techniques, showing their suitability for a wide range of existing conformal prediction algorithms. We choose Conformalized Quantile Regression (CQR, Romano et al. [2019b]) to propose Fast Feature Conformalized Quantile Regression (FFCQR). The fundamental concept is similar to FFCP Algorithm 2, where calibration steps are performed within the gradient information. FFCQR algorithm is proposed in Algorithm 3.

Algorithm 3 Fast Feature Conformalized Quantile Regression (FFCQR)

Input: Confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\mathcal{I}}$, test point X';

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tra}} \triangleq (X_i, Y_i)_{i\in\mathcal{I}_{\text{tra}}}$ together with a calibration fold $\mathcal{D}_{\text{cal}} \triangleq (X_i, Y_i)_{i\in\mathcal{I}_{\text{cal}}}$;

2: Train a base machine learning model $f^{\text{lo}} = g^{\text{lo}} \circ h(\cdot)$ and $f^{\text{hi}} = g^{\text{hi}} \circ h(\cdot)$ using $\mathcal{D}_{\text{tra}}$ to estimate the quantile of response Y_i, which returns $[f^{\text{lo}}(X_i), f^{\text{hi}}(X_i)]$;

3: For each $i \in \mathcal{I}_{\text{cal}}$, calculate the non-conformity score $\tilde{R}_i^{\text{lo}} = (f^{\text{lo}}(X_i) - Y_i)/\|\nabla g^{\text{lo}}(\hat{v}_i)\|$ and $\tilde{R}_i^{\text{hi}} = (Y_i - f^{\text{hi}}(X_i))/\|\nabla g^{\text{hi}}(\hat{v}_i)\|$, where $\nabla g(\cdot)$ denote the gradient of $g(\cdot)$ on the feature $\hat{v}_i \triangleq h(X_i)$, namely $\nabla g^{\text{lo}}(\hat{v}_i) = \frac{\mathrm{d}g^{\text{lo}}\circ h(X_i)}{\mathrm{d}h(X_i)}$ and $\nabla g^{\text{hi}}(\hat{v}_i) = \frac{\mathrm{d}g^{\text{hi}}\circ h(X_i)}{\mathrm{d}h(X_i)}$

4: Calculate the $(1-\alpha/2)$-th quantile $Q_{1-\alpha/2}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{cal}}|+1}\sum_{i\in\mathcal{I}_{\text{cal}}}\delta_{\tilde{R}_i} + \delta_\infty$, where $\tilde{R}_i = \max\left\{\tilde{R}_i^{\text{lo}}, \tilde{R}_i^{\text{hi}}\right\}$

Output: $\mathcal{C}_{1-\alpha/2}^{\text{ffcqr}}(X') = \left[f^{\text{lo}}(X') - \|\nabla g^{\text{lo}}(\hat{v}')\| \cdot Q_{1-\alpha/2}, f^{\text{hi}}(X') + \|\nabla g^{\text{hi}}(\hat{v}')\| \cdot Q_{1-\alpha/2}\right]$, where $\hat{v}' = h(X')$.

We summarize run time in Table 8 and the experiments result in Table 9 (meps19), Table 10 (com), and Table 11 (bike). FFCQR reduces runtime compared to FCQR, while achieving better efficiency compared to CQR.

Furthermore, we have observed that as the values of $[\alpha, 1-\alpha]$ used by the neural networks in all CQR methods (CQR, FCQR and FFCQR) become increasingly closer in the training process (The level difference between [0.1, 0.9] is 0.8, while the level difference between [0.49, 0.51] is 0.02, with the difference gradually decreasing), the band length returned by FFCQR gradually narrows. This implies that our method holds an advantage on returned band length when the narrower neural network confidence level.

Table 8: Time Comparison among CQR, FCQR and FFCQR. For quantile regression tasks, FFCQR also demonstrates more efficient performance. The last column represents the speed improvement factor of FFCQR compared to FCQR. The time unit is in seconds.

DATASET	CQR	FCQR	FFCQR	FASTER
SYNTHETIC	$0.0125_{\pm 0.0062}$	$0.3237_{\pm 0.0152}$	$0.0742_{\pm 0.0091}$	4X
COM	$0.0045_{\pm 0.0015}$	$0.2730_{\pm 0.1088}$	$0.0210_{\pm 0.0011}$	13X
FB1	$0.0446_{\pm 0.0157}$	$1.7276_{\pm 0.1389}$	$0.2532_{\pm 0.0166}$	7X
FB2	$0.0812_{\pm 0.0187}$	$3.9967_{\pm 0.7330}$	$0.0617_{\pm 0.0123}$	65X
MEPS19	$0.0187_{\pm 0.0018}$	$0.7671_{\pm 0.0438}$	$0.1189_{\pm 0.0048}$	6X
MEPS20	$0.0438_{\pm 0.0079}$	$1.1876_{\pm 0.2206}$	$0.1505_{\pm 0.0138}$	8X
MEPS21	$0.0187_{\pm 0.0027}$	$0.8004_{\pm 0.0657}$	$0.1120_{\pm 0.0053}$	7X
STAR	$0.0047_{\pm 0.0009}$	$0.2352_{\pm 0.0419}$	$0.0214_{\pm 0.0005}$	11X
BIO	$0.0774_{\pm 0.0541}$	$6.9365_{\pm 4.5494}$	$0.6473_{\pm 0.4879}$	11X
BLOG	$0.1121_{\pm 0.0153}$	$1.9591_{\pm 0.1346}$	$0.3941_{\pm 0.0618}$	5X
BIKE	$0.0138_{\pm 0.0045}$	$1.8528_{\pm 2.3969}$	$0.2382_{\pm 0.3261}$	6X

Table 9: Coverage and Band Length at Different Net Confidence Levels Used By the Neural Networks in CQR methods with *Meps19* detaset. FFCQR yields shorter band lengths compared to CQR.

Confidence Levels		[0.1, 0.9]		[0.2, 0.8]		[0.3, 0.7]		[0.4, 0.6]		[0.49, 0.51]	
Metrics		Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
Split CQR		90.28±0.47	2.43±0.11	90.19±0.46	2.58±0.45	90.63±0.32	2.93±0.51	90.48±0.42	3.47±0.16	90.44±0.36	3.48±0.08
FCQR		91.32±0.37	1.50±0.37	90.26±0.33	2.61±2.01	90.45±0.54	2.30±2.38	90.58±0.33	6.11±0.89	90.47±0.45	4.59±1.73
FFCQR	layer0	90.29±0.60	2.61±0.13	90.22±0.26	2.58±0.54	90.31±0.43	3.03±0.90	89.95±0.29	5.30±0.57	89.84±0.12	5.96±1.26
	layer1	90.29±0.57	2.56±0.13	90.10±0.33	2.49±0.52	90.34±0.46	2.92±0.85	90.02±0.33	5.07±0.50	89.88±0.24	5.71±1.22
	layer2	90.14±0.60	2.34±0.12	90.24±0.65	2.22±0.32	90.34±0.43	2.60±0.68	89.96±0.40	4.10±0.22	89.97±0.33	4.76±1.17
	layer3	90.21±0.45	2.18±0.12	90.14±0.42	2.10±0.19	90.35±0.36	2.34±0.19	90.28±0.33	2.76±0.15	89.86±0.49	3.16±0.62
	layer4	90.28±0.47	2.43±0.11	90.19±0.46	2.58±0.45	90.63±0.32	2.93±0.51	90.48±0.42	3.47±0.16	90.44±0.36	3.48±0.08

Table 10: Coverage and Band Length at Different Net Confidence Levels Used By the Neural Networks in CQR methods with *com* dataset. FFCQR yields shorter band lengths compared to CQR.

Confidence Levels		[0.1, 0.9]		[0.2, 0.8]		[0.3, 0.7]		[0.4, 0.6]		[0.49, 0.51]	
Metrics		Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
Split CQR		89.87±1.68	1.57±0.12	90.13±0.89	1.71±0.18	89.87±1.06	1.74±0.16	89.27±0.94	2.07±0.55	89.57±0.49	1.99±0.12
FCQR		90.83±1.53	1.19±0.19	90.43±1.32	0.49±0.38	90.23±1.13	0.37±0.06	90.18±1.77	0.20±0.05	89.47±0.85	0.23±0.07
FFCQR	layer0	88.92±2.78	1.62±0.12	89.62±1.75	1.67±0.07	91.53±0.97	1.62±0.12	89.77±1.64	1.80±0.27	89.82±1.07	1.76±0.11
	layer1	88.67±2.40	1.59±0.12	89.57±1.03	1.64±0.08	90.58±1.21	1.57±0.13	89.82±1.75	1.78±0.33	89.12±1.40	1.74±0.09
	layer2	89.77±2.14	1.58±0.12	89.92±1.98	1.63±0.12	90.53±0.43	1.64±0.14	89.67±1.28	1.89±0.38	88.77±0.75	1.78±0.11
	layer3	90.08±2.28	1.58±0.12	89.92±1.22	1.67±0.15	90.33±0.86	1.73±0.13	89.62±0.83	2.03±0.54	89.27±0.66	1.93±0.11
	layer4	89.87±1.68	1.57±0.12	90.13±0.89	1.71±0.18	89.87±1.06	1.74±0.16	89.27±0.94	2.07±0.55	89.57±0.49	1.99±0.12

B.6 Group coverage

Group coverage is represented by the conditional probability $\mathbb{P}(Y \in \mathcal{C}(X)|X)$. The test dataset was categorized into three groups by splitting response Y based on the lower and upper tertiles, and we have reported the minimum coverage for each group.

We present our results in two parts: (a) we present the group coverage provided by Split CP, FCP, FFCP, detailed in Table 14 and (b) the group coverage provided by LCP and FFLCP, as shown in Table 13.

Analyzing the experimental results, we believe that the group coverage achieved through gradient-level techniques in FFCP reflects an improvement over Split CP, albeit with moderate overall performance. We note that the group coverage of gradient-level conformal prediction is contingent upon its Split version. That is, when the Split version demonstrates satisfying group coverage, the gradient-level version tends to mirror this result. Thus, despite FFCP outperforming Split CP, the overall performance is still considered average.

LCP, developed specifically to enhance group coverage, inherently achieves higher coverage. Experimental results further reveal that FFLCP surpasses LCP, demonstrating the superiority of our gradient-level techniques.

Algorithm 4 Fast Feature Localized Conformal Prediction (FFLCP)

Input: Confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\mathcal{I}}$, tesing point X', localizer $D(X, Y)$

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tra}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{tra}}}$ and a calibration fold $\mathcal{D}_{\text{cal}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$;

2: Train a base neural network with training fold $f(\cdot) = g \circ h(\cdot)$ with training fold $\mathcal{D}_{\text{tra}}$;

3: For each $i \in \mathcal{I}_{\text{cal}}$, calculate the non-conformity score $\tilde{R}_i = |Y_i - f(X_i)|/\|\nabla g(\hat{v}_i)\|$, where $\nabla g(\hat{v}_i)$ denotes the gradient of $g(\cdot)$ on the feature $\hat{v}_i \triangleq h(X_i)$, namely $\nabla g(\hat{v}_i) = \frac{\mathrm{d}g\circ h(X_i)}{\mathrm{d}h(X_i)}$;

4: Calculate the distance $D_i \triangleq D(X', X_i)$, $d_i^D := \frac{D_i}{\sum_{i\in\mathcal{I}_{\text{cal}}} D_i}$ and $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\sum_{i\in\mathcal{I}_{\text{cal}}} d_i^D \delta_{\tilde{R}_i} + \delta_\infty$;

Output: $\mathcal{C}_{1-\alpha}^{\text{fflcp}}(X') = [f(X') - \|\nabla g(\hat{v}')\|Q_{1-\alpha}, f(X') + \|\nabla g(\hat{v}')\|Q_{1-\alpha}]$, where $\hat{v}' = h(X')$.

Table 11: Coverage and Band Length at Different Net Confidence Levels Used By the Neural Networks in CQR methods with *bike* dataset. FFCQR yields shorter band lengths compared to CQR.

Confidence Levels		[0.1, 0.9]		[0.2, 0.8]		[0.3, 0.7]		[0.4, 0.6]		[0.49, 0.51]	
Metrics		Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
Split CQR		89.38±0.73	0.82±0.07	89.99±0.69	0.73±0.03	89.63±0.84	0.77±0.03	90.25±0.62	0.84±0.02	89.72±0.51	0.96±0.08
FCQR		90.25±0.67	0.58±0.15	90.14±0.63	0.65±0.13	89.77±0.76	0.71±0.18	89.93±0.35	0.82±0.07	89.98±0.97	0.74±0.08
FFCQR	layer0	89.91±0.38	0.91±0.07	89.84±0.44	0.97±0.02	89.42±0.83	1.20±0.04	89.75±0.54	1.64±0.13	89.61±0.59	1.79±0.08
	layer1	89.57±0.33	0.90±0.07	89.83±0.25	0.90±0.05	89.44±0.42	1.04±0.07	89.72±0.43	1.25±0.05	89.62±0.73	1.31±0.06
	layer2	89.73±0.29	0.87±0.07	89.70±0.27	0.79±0.03	89.63±0.69	0.83±0.03	89.14±0.42	0.92±0.04	89.44±0.41	0.98±0.04
	layer3	89.49±0.34	0.84±0.06	89.62±0.48	0.69±0.02	89.58±0.74	0.69±0.02	89.86±0.37	0.70±0.01	89.57±0.88	0.78±0.07
	layer4	89.38±0.73	0.82±0.07	89.99±0.69	0.73±0.03	89.63±0.84	0.77±0.03	90.25±0.62	0.84±0.02	89.72±0.51	0.96±0.08

Table 12: Coverage and Band Length at Different Net Confidence Levels Used By the Neural Networks in CQR methods with *bio* dataset. FFCQR yields shorter band lengths compared to CQR.

Confidence Levels		[0.1, 0.9]		[0.2, 0.8]		[0.3, 0.7]		[0.4, 0.6]		[0.49, 0.51]	
Metrics		Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
Split CQR		89.89±0.41	1.42±0.02	89.84±0.27	1.45±0.02	89.87±0.27	1.61±0.02	90.07±0.31	1.86±0.03	90.16±0.40	2.00±0.03
FCQR		90.18±0.35	0.95±0.50	90.45±0.45	2.09±0.41	90.16±0.48	1.84±0.43	90.25±0.46	2.37±0.76	90.21±0.46	2.02±0.34
FFCQR	layer0	89.74±0.32	1.47±0.01	89.98±0.22	1.56±0.04	89.89±0.25	1.73±0.04	89.87±0.24	2.22±0.15	89.64±0.20	2.55±0.06
	layer1	89.77±0.33	1.45±0.01	89.99±0.21	1.48±0.03	89.92±0.37	1.59±0.03	89.92±0.21	1.99±0.12	89.69±0.28	2.21±0.04
	layer2	89.77±0.40	1.43±0.02	90.01±0.23	1.41±0.01	90.02±0.32	1.49±0.03	90.01±0.49	1.76±0.11	89.79±0.35	1.94±0.07
	layer3	89.75±0.41	1.41±0.02	89.98±0.34	1.38±0.02	89.93±0.41	1.47±0.01	90.07±0.12	1.68±0.04	89.97±0.34	1.78±0.02
	layer4	89.89±0.41	1.42±0.02	89.84±0.27	1.45±0.02	89.87±0.27	1.61±0.02	90.07±0.31	1.86±0.03	90.16±0.40	2.00±0.03

B.7 FFRAPS

In this section, we show how to deploy gradient-level techniques in FFCP in classification problems. The basic ideas follow Algorithm 5.

Comparing to the experimental part of RAPS, our core adjustments are as follows:

(a) During the calibration process, for the model's output of sorted scores s, we divide each element by the magnitude of its corresponding gradient: $s+\delta \cdot \|\nabla g(v)\|$. Here, δ is an adjustable hyper-parameter that can be tuned to optimize the performance of the model based on the specific characteristics of the data and the problem at hand.

(b) In the stage of calculating the returned set, we multiply the generalized inverse quantile τ by the magnitude of the gradient of the corresponding test data: $s' + \delta \cdot \|\nabla g(v')\|$

We summarize the experiment results in Table 15, where we adhere to the statistical methodologies of RAPS as described in Angelopoulos et al. [2020].

Algorithm 5 Fast Feature Regularized Adaptive Prediction Sets (FFRAPS)

Input: Confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\mathcal{I}}$, tesing point X', and ground-truth label $y \in \{0, 1, ..., K\}^n$ for $X \in \mathcal{D}$ and X'; regularization hyperparameters k_{reg}, δ and λ;

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tra}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{tra}}}$ and a calibration fold $\mathcal{D}_{\text{cal}} \triangleq \{(X_i, Y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$;

2: Train a base neural network with training fold $f(\cdot) = g \circ h(\cdot)$ with training fold $\mathcal{D}_{\text{tra}}$;

3: For each $i \in \mathcal{I}_{\text{cal}}$, $L_i \leftarrow j$ such that $I_{i,j} = y_i$, where I represents the associated permutation of index. Calculate generalized inverse quantile conformity score $E_i = \Sigma_{j=1}^{L_i} s_{i,j} + \|\nabla g(\hat{v}_i)\| \cdot \delta + \lambda(L_i - k_{reg})^+$, where $\nabla g(\hat{v}_i)$ denotes the gradient of $g(\cdot)$ on the feature $\hat{v}_i \triangleq h(X_i)$, namely $\nabla g(\hat{v}_i) = \frac{\mathrm{d}g\circ h(X_i)}{\mathrm{d}h(X_i)}$, where $s \triangleq \text{sort}f(X)$ represents the sorted scores. Calculate $\hat{\tau}_{ccal} \leftarrow \lceil (1-\alpha)(1+n) \rceil$ largest value in $\{E_i\}_{i=1}^n$

4: Calculate $L \leftarrow |\ \{j \in \mathcal{Y} \ : \ \Sigma_{i=1}^{j} s'_i + \|\nabla g(\hat{v}'_i)\| \cdot \delta + \lambda(j - k_{reg})^+ \leq \hat{\tau}_{ccal*}\}\ | + 1$, where $\hat{v}' = h(X')$ and $s' = \text{sort}f(X')$;

Output: $\mathcal{C}_{1-\alpha}^{\text{FFRAPS}}(X') = \{I_1, ...I_L\}$

Table 13: Comparison of LCP and FFLCP in group coverage. We divide the datasets into three groups based on the size of Y, and calculate the coverage for each group, returning the maximum coverage. FFLCP shows the results for the 5-layer neural network.

METHOD	LCP	FFLCP				
DATASET	COVERAGE	LAYER0	LAYER1	LAYER2	LAYER3	LAYER4
SYNTHETIC	87.02±1.00	86.93±0.78	86.57±0.88	85.43±1.24	**87.11**±1.76	87.02±1.00
COM	80.33±3.24	**81.84**±2.52	79.56±3.34	77.42±4.12	79.41±2.88	80.33±3.24
FB1	52.51±1.76	**78.61**±0.91	76.39±1.21	67.16±1.61	57.82±1.88	52.51±1.76
FB2	54.33±1.75	**75.86**±0.83	75.44±0.91	70.45±0.99	60.18±1.73	54.33±1.75
MEPS19	67.35±1.21	**68.19**±2.31	66.44±2.01	64.94±1.53	67.14±1.24	67.35±1.21
MEPS20	65.49±1.64	**69.30**±1.09	68.80±1.55	65.14±1.44	65.47±1.99	65.49±1.64
MEPS21	66.38±0.95	67.82±1.10	**67.96**±1.21	66.21±1.33	65.54±1.02	66.38±0.95
STAR	77.20±3.97	**79.69**±2.88	79.28±1.72	77.47±5.21	77.33±4.12	77.20±3.97
BIO	81.10±0.61	86.33±0.51	86.06±0.50	**86.78**±0.57	83.26±0.71	81.10±0.61
BLOG	48.99±1.01	**61.01**±0.82	55.10±1.12	46.01±0.65	46.88±0.81	48.99±1.01
BIKE	77.61±1.52	81.02±1.73	82.42±2.08	82.97±1.29	**84.41**±1.71	77.61±1.52

Table 14: Comparison of Split CP, FCP and FFCP in group coverage.

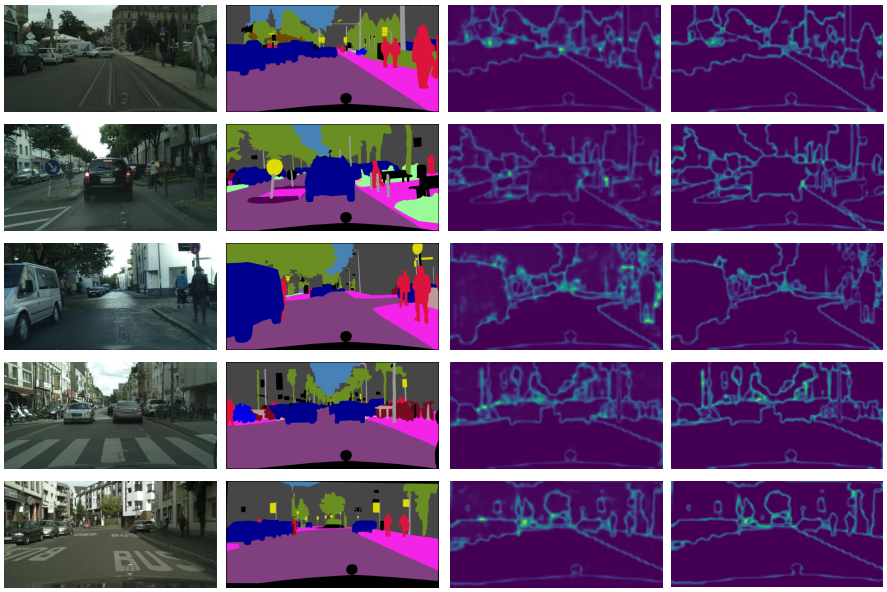
METHOD	SPLIT CP	FCP	FFCP				
DATASET	COVERAGE	COVERAGE	LAYER0	LAYER1	LAYER2	LAYER3	LAYER4
SYNTHETIC	87.08±1.03	87.92±1.08	86.96±0.81	86.63±0.79	85.64±1.13	**88.46**±1.44	87.08±1.03
COM	79.41±3.12	79.57±2.96	**82.00**±3.18	79.41±3.62	78.64±4.35	78.65±3.62	79.41±3.12
FB1	56.69±1.35	57.34±1.12	**79.20**±0.95	76.75±1.42	68.09±1.76	59.33±1.91	56.69±1.35
FB2	57.98±1.28	58.72±0.87	**76.27**±0.92	75.64±0.91	70.86±0.89	62.43±1.15	57.98±1.28
MEPS19	73.78±1.08	**73.82**±0.91	70.90±2.29	70.51±2.28	72.09±1.25	73.53±1.00	73.78±1.08
MEPS20	72.21±1.47	**72.33**±1.46	70.42±0.88	70.13±1.42	69.51±0.79	71.17±2.01	72.21±1.47
MEPS21	71.38±0.20	**72.02**±0.70	69.40±1.61	69.83±1.44	69.81±1.68	70.85±0.82	71.39±0.20
STAR	**83.45**±3.09	83.17±3.47	82.89±1.51	81.22±2.55	81.22±3.60	83.03±2.07	**83.45**±3.09
BIO	81.00±0.61	84.45±0.88	87.31±0.27	87.27±0.46	**88.31**±0.72	84.20±0.70	81.00±0.61
BLOG	58.32±0.90	60.43±1.46	**65.21**±0.58	59.03±1.03	54.55±0.77	55.76±1.26	58.32±0.90
BIKE	77.55±1.40	86.25±0.87	**95.36**±1.32	94.23±1.40	95.06±1.06	84.65±1.85	77.55±1.40

B.8 Additional Experiment Results

This section provides more experiment results. Additional visual results for the segmentation problem are also presented in Figure 5.

Table 15: Comparison of FFRAPS with the state-of-the-art method RAPS on Imagenet-Val. The FFRAPS method outperforms RAPS in most datasets.

METHOD	ACCURACY		COVERAGE		LENGTH	
MODEL	TOP-1	TOP-5	RAPS	FFRAPS	RAPS	FFRAPS
RESNEXT101	0.793±0.001	0.945±0.001	0.908±0.002	0.907±0.002	2.012±0.035	**2.006**±0.039
RESNET152	0.784±0.001	0.941±0.001	0.909±0.003	0.907±0.003	2.144±0.034	**2.128**±0.058
RESNET101	0.774±0.001	0.935±0.001	0.906±0.004	0.906±0.003	2.348±0.151	**2.256**±0.037
RESNET50	0.761±0.001	0.929±0.001	0.907±0.004	0.907±0.003	**2.560**±0.104	2.594±0.069
RESNET18	0.698±0.001	0.891±0.001	0.906±0.003	0.903±0.003	4.560±0.147	**4.434**±0.168
DENSENET161	0.772±0.001	0.936±0.001	0.907±0.003	0.907±0.002	2.374±0.083	**2.328**±0.056
VGG16	0.716±0.001	0.904±0.001	0.904±0.002	0.902±0.002	3.566±0.098	**3.521**±0.065
INCEPTION	0.696±0.001	0.887±0.001	0.903±0.003	0.903±0.002	5.410±0.350	**5.407**±0.133
SHUFFLENET	0.694±0.001	0.883±0.001	0.902±0.001	0.901±0.002	5.001±0.121	**4.971**±0.073



(a) Image (b) Ground Truth Label (c) FCP Length (d) FFCP Length

Figure 5: Additional visualization results in segmentation task.

B.9 Additional Experiments on Full CP

We further validate the generality of our proposed approach by applying it to the CV+ (Barber et al. [2021]) framework. Experiments were conducted on a synthetic multi-class classification dataset generated using scikit-learn. A total of 1,200 samples were created, with 1,000 used for training and 200 for testing. The dataset contained 10 features, including 3 informative and 2 redundant ones, distributed across 8 distinct classes, each consisting of a single cluster. Our method achieves higher efficiency while maintaining comparable coverage to the standard CV+ approach.

Table 16: Comparison of FFCV+ with CV+ on synthetic dataset.

NUM	CLASS = 3		CLASS = 5		CLASS = 8	
METHOD	COVERAGE	SIZE	COVERAGE	SIZE	COVERAGE	SIZE
CV+	90.50±0.01	1.23±0.01	90.00±0.01	1.57±0.01	92.00±0.01	2.13±0.01
FFCV+	91.00±0.01	1.20±0.01	92.50±0.01	1.55±0.01	91.50±0.01	2.11±0.01

B.10 Additional Experiments on Self-Supervised CP

SSCP (Seedat et al. [2023]) is a method that leverages self-supervised signals to enhance the adaptability and efficiency of conformal prediction intervals. As shown in Table 1, however, SSCP introduces substantially higher computational overhead due to the need to train two additional networks, making it approximately 2 to 50 times slower than FFCP. Across all settings, FFCP achieves shorter prediction bands, which may be attributed to SSCP's stronger dependence on the base model's predictive quality and the complexity of its auxiliary network training.

Table 17: Results on FFCP and SSCP

METHOD	SSCP		FFCP	
DATASET	TIME	LENGTH	TIME	LENGTH
SYNTHETIC	7.18±0.99	**0.25**±0.02	0.15±0.02	0.18±0.01
COM	1.33±0.16	2.52±0.14	0.03±0.01	**1.84**±0.18
MEPS19	10.86±0.76	5.32±0.33	0.14±0.01	**3.13**±0.30
STAR	1.96±0.02	0.67±0.06	0.67±0.06	**0.21**±0.01
BIO	13.49±3.81	**1.53**±0.03	0.39±0.04	1.66±0.02
BIKE	5.92±0.03	**0.73**±0.04	0.09±0.01	0.63±0.03

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 state the key theoretical and empirical contributions, consistent with the main results.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the main limitations, including assumptions on data distribution and computational scalability, in the paper.

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: All assumptions are clearly stated alongside the theorems, with complete proofs included in the appendix.

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: Experimental settings and evaluation procedures are described in detail in the main text and appendix.

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: An anonymized link to the code and data is provided in the supplemental material along with usage instructions.

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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Full details about training, testing, hyperparameters, and optimizers are included in the main text and appendix.

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 report standard deviations across multiple runs to indicate the statistical significance of the results.

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: Details on hardware specifications, runtime, and total compute budget are provided in the appendix.

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 adheres to the NeurIPS Code of Ethics and presents no ethical concerns.

10. **Broader impacts**

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

    Answer: [NA]

    Justification: This work is theoretical and does not directly relate to applications with societal impact.

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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve any models or datasets with potential risk of misuse.

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: All external assets are properly cited with license terms stated in the appendix or main text.

13. **New assets**

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

    Answer: [NA]

    Justification: No new assets are introduced in this work.

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: [NA]

    Justification: The research does not involve any human subjects or crowdsourcing.

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: [NA]

    Justification: The study does not involve human subjects and thus does not require IRB approval.

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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not use LLMs as part of its core methodology.