Occupancy-based Quantile Risk Control

Zihao Shi[1]*, **Huajun Xi**[2], **Bingyi Jing**[3,4], **Hongxin Wei**[1]†
[1]Southern University of Science and Technology
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]The Chinese University of Hong Kong, Shenzhen [4]Shenzhen Loop Area Institute

Abstract

Conformal risk control is an emerging framework for the safe deployment of machine learning models with finite-sample guarantees. To accommodate a broader class of risk notions, quantile risk control extends this framework to quantile-based risk measures. However, existing methods either suffer from excessive conservatism or lack rigorous finite-sample guarantees. To address these limitations, we introduce Occupancy-based Quantile Risk Control (OQRC), a novel method that provides tight risk control bounds with finite-sample validity. Our key idea is to formulate risk control as a finite-occupancy problem by partitioning the loss space with the ordered calibration losses. Specifically, we estimate the distribution of test losses across the resulting bins and upper-bound the risk by the maximum loss attained within each bin. We then select the parameter λ such that this upper bound does not exceed a predefined threshold α with high probability $1-\delta$. Theoretically, we establish a finite-sample guarantee showing that OQRC yields tight risk control bounds that converge to the optimal bounds at a provable rate of $\mathcal{O}_p(n^{-1/2})$. Extensive experiments demonstrate the effectiveness of our method, reducing the risk gap by up to 78.64% on common benchmarks.

1 Introduction

Machine learning (ML) models are increasingly deployed in risk-sensitive domains, such as medical diagnostics [1], finance [2], and autonomous driving [3]. Despite their high accuracy, the black-box nature of ML models makes them untrustworthy for users due to the uncontrollable risk of predictions, e.g., imaging-related errors in radiology [4] and misdiagnoses in healthcare [5]. To this end, Conformal Risk Control (CRC) [6] introduces a model-agnostic framework that guarantees the prediction risk remains below a predefined threshold. With a monotone loss function, the framework can be applied to control the false-negative rate in classification tasks [7] or to improve the factuality of outputs generated by large language models [8]. By offering such flexible risk guarantees, it presents a rigorous and feasible solution toward the trustworthy deployment of black-box ML models.

While elegant and simple, the framework of conformal risk control is limited to controlling the expected loss. To address broader risk notions, Quantile Risk Control (QRC) [9] extends this framework by providing guarantees on quantile-based risk measures, such as value-at-risk [10] and conditional value-at-risk [11]. Despite these advancements, this method tends to overestimate prediction risk, leading to conservative results. To alleviate this issue, recent work [12] leverages L-statistics [13] to derive tight risk control bounds. However, their theoretical results rely on the asymptotic normality of L-statistics and offer no guarantees for finite-sample regimes. This motivates us to establish a non-conservative risk control method with finite-sample guarantees.

*Work done while working at SUSTech as a visiting scholar.
†Correspondence to: Hongxin Wei <weihx@sustech.edu.cn>

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

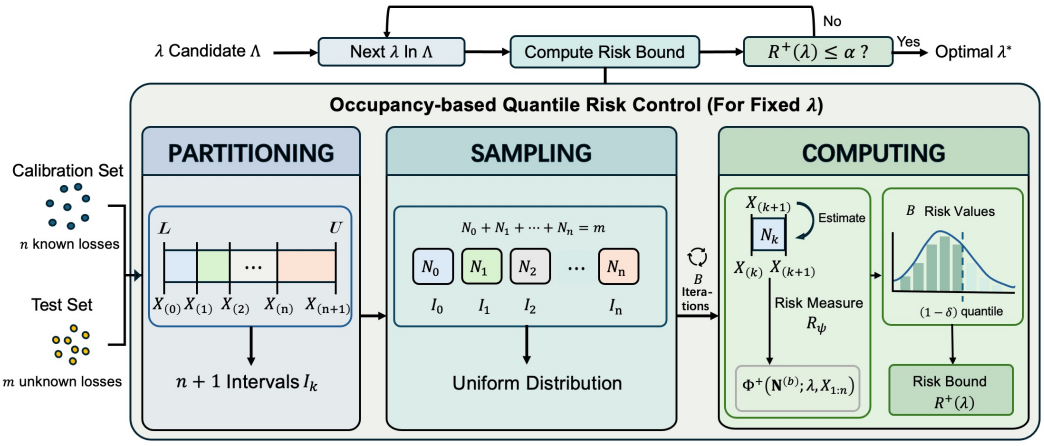


Figure 1: **Overview of Occupancy-based Quantile Risk Control.** Given a risk level α and confidence $1-\delta$, our OQRC implements a sequential procedure that calibrates a parameter λ^* of the ML model such that the quantile risk R_{test} satisfies $\mathbb{P}(R_{\text{test}}(\lambda^*) \le \alpha) \ge 1-\delta$.

In this work, we introduce Occupancy-based Quantile Risk Control (OQRC), a novel risk control method that provides tight risk control bounds with finite-sample guarantees (see Figure 1). Our key idea is to formulate risk control as a finite-occupancy problem by partitioning the loss space with the ordered calibration losses. Specifically, we estimate the distribution of test losses across the resulting intervals and upper-bound the risk by the maximum loss attained within each interval. We then select the parameter λ such that this upper bound does not exceed a predefined threshold α with high probability $1-\delta$. Theoretically, we establish a finite-sample guarantee showing that OQRC yields tight risk control bounds that converge to the optimal bounds at a rate of $\mathcal{O}_p(n^{-1/2})$.

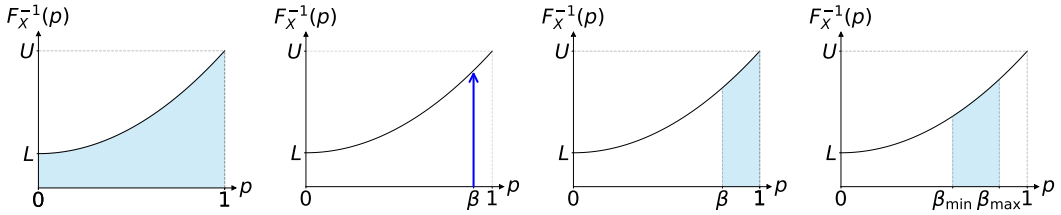
To validate the effectiveness of our method, we conduct extensive experiments on seven real-world datasets across three tasks: tumor segmentation [14, 15], multi-label image classification [16], and text emotion recognition [17]. For each task, we encompass three risk measures: VaR, CVaR, and VaR-Interval. To assess the conservatism of risk control, we use a new metric – *RiskGap*, defined as the empirical average of the absolute difference between the empirical and target risks. Empirical results show that OQRC significantly reduces the RiskGap compared to prior methods. For instance, when using false-negative proportion as the loss function for multi-label image classification, our method reduces the RiskGap by 78.64% compared to alternative approaches. Moreover, OQRC exhibits consistent performance across various hyperparameter configurations and model architectures, highlighting its robustness. Finally, we apply OQRC to unconditional control [6], and the results indicate that OQRC successfully bounds the risk below α across different risk measures. In summary, these results demonstrate the effectiveness, robustness, and superiority of our method. Our code is available at https://github.com/ml-stat-Sustech/OQRC.

We summarize our contributions as follows:

- We introduce Occupancy-based Quantile Risk Control (OQRC), a simple and effective method that reformulates risk control as a finite occupancy problem. We demonstrate that OQRC is also applicable to unconditional risk control, highlighting its versatility.
- Theoretically, we establish rigorous guarantees demonstrating that OQRC achieves finite-sample quantile risk control. In addition, we prove the tightness of OQRC by analyzing the convergence rate of the corresponding risk control bounds to the optimal bounds.
- Through extensive empirical validation, we demonstrate that OQRC significantly narrows the risk gap compared to alternative methods. Our further evaluations validate the robustness of OQRC across diverse hyperparameter configurations and model architectures.

2 Preliminaries

Problem Setup. We consider a risk control problem. Let $h : \mathcal{Z} \to \widehat{\mathcal{Y}}$ be a black-box predictor that maps an input space $\mathcal{Z} \subset \mathbb{R}^d$ to a prediction space $\widehat{\mathcal{Y}}$. We assume a loss function $\ell_\lambda : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [L, U]$



(a) Expected loss (b) β-VaR (c) β-CVaR (d) VaR-Interval

Figure 2: **Examples of quantile-based risk measures.** The horizontal axis represents the quantile, and the vertical axis denotes the loss value. Expected loss, β-CVaR, and $[\beta_{\min}, \beta_{\max}]$-VaR-Interval indicate the average loss within the sky-blue region, while β-VaR indicates the loss at the β-quantile.

that quantifies the quality of a prediction $\hat{Y}$ with respect to the target output Y. $\lambda \in \Lambda$ is a tunable parameter within the ML model, used to control the risk level of predictions. Given data $(Z, Y) \in \mathcal{D}$, we define the random variable $X \coloneqq \ell_\lambda(h(Z), Y)$ to be the loss induced by h on $\mathcal{D}$. Recall that the cumulative distribution function (CDF) of a random variable X is defined as $F_X(x) \coloneqq \mathbb{P}(X \leq x)$. The goal of the risk control problem is to choose the parameter λ to ensure that

$$\mathbb{P}\big(R(F_X; \lambda) \leq \alpha\big) \geq 1 - \delta. \tag{1}$$

where $R(\cdot; \lambda) \in \mathcal{R}$ is a risk measure on F_X, α is a predefined risk level, and δ is a confidence level.

To select a valid λ, a standard practice is to hold out a calibration set, compute the corresponding losses $X_{1:n} = \{X_1, ..., X_n\}$, and subsequently utilize these observed losses to calculate λ. Following the previous work [6, 18], we have the following mild assumptions:

Assumption 2.1. The observed losses on the calibration set $X_{1:n}$ and the unknown losses on the test set $X^{\text{test}}_{1:m} = \{X^{\text{test}}_1, \ldots, X^{\text{test}}_m\}$ are independent and identically distributed (i.i.d.).

Assumption 2.2. Losses in $X_{1:n} \cup X^{\text{test}}_{1:m}$ are non-increasing and right-continuous across λ.

Notably, a strictly minimized $R(F_X; \lambda)$ is not optimal in a risk control problem, since an extremely low risk often incurs a sacrifice of valid information [6]. Consider a scenario where a model identifies tumor pixels in medical images and the risk represents the proportion of the missed tumor pixels. Selecting every pixel in the image minimizes this risk while rendering the output entirely uninformative. Therefore, provided the risk control objective is met, a value of $R(F_X; \lambda)$ closer to the target level α is preferred because it leads to predictions containing more useful information.

Quantile Risk Control. We focus on a general class of risk measures identified as quantile-based risk measures (QBRMs). The quantile function is defined as $F_X^{-1}(p) \coloneqq \inf\{x : F_X(x) \geq p\}$.

Definition 2.3 (Quantile-based Risk Measure)**.** Let $\psi(p)$ denote a weighting function such that $\psi(p) \geq 0$ and $\int_0^1 \psi(p)\, dp = 1$. The quantile-based risk measure parameterized by ψ is expressed as

$$R_\psi(F_X; \lambda) \coloneqq \int_0^1 \psi(p) F_X^{-1}(p)\, dp.$$

Figure 2 illustrates several representative examples of QBRMs. Specifically, the expected loss corresponds to $\psi(p) = 1$. For the β-VaR, we define $\psi(p) = \delta_\beta(p)$, where δ_β denotes the Dirac delta function centered at β. For the β-CVaR, we define $\psi(p) = \mathbf{1}\{p \geq \beta\}\frac{1}{1-\beta}$. Finally, the $[\beta_{\min}, \beta_{\max}]$-VaR-Interval is formulated using $\psi(p) = \mathbf{1}\{\beta_{\min} \leq p \leq \beta_{\max}\}\frac{1}{\beta_{\max} - \beta_{\min}}$.

Prior studies have investigated the rigorous, distribution-free control of quantile-based risk measures. Snell et al. [9] achieve quantile risk control by constructing a lower confidence bound F_X^- for F_X through a one-sided test statistic. However, this approach often tends to overestimate prediction risk, yielding conservative risk control bounds. To alleviate this issue, Chen et al. [12] control the risk via L-statistics. They use L-statistics to estimate the empirical distribution $\widehat{F}_n$ of F_X, and then control the risk by adding a correction term to the corresponding risk measures of $\widehat{F}_n$, thereby deriving tighter risk control bounds than before. Nevertheless, the theoretical guarantees of this method hold only asymptotically, making the approach potentially fail in finite-sample settings. These limitations motivate us to develop a method that provides both tight bounds and finite-sample guarantees.

3 Method

We propose a novel method for quantile risk control termed *Occupancy-based Quantile Risk Control* (OQRC). The key idea is to reformulate risk control as a finite occupancy problem [19] by partitioning the loss space using calibration data. We use these resulting partitioned bins to construct an upper confidence bound (UCB) $R^+(\lambda)$ such that for every λ, we have

$$\mathbb{P}(R_\psi(F_X;\lambda) \le R^+(\lambda)) \ge 1-\delta. \tag{2}$$

Subsequently, we achieve valid risk control (1) by identifying $\lambda^* := \inf\{\lambda \in \Lambda : R^+(\lambda) \le \alpha\}$.

In the following, we present the details of this procedure. We define the empirical CDF on test set as $\widehat{F}_m(x) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{X_i^{\text{test}} \le x\}$. The risk on the test set is then given by $R_{\text{test}}(\lambda) = R_\psi(\widehat{F}_m;\lambda)$.

3.1 Upper Confidence Bound via Partitioned Occupancy

We first partition the loss space using the calibration losses to construct the bins for the occupancy problem. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the order statistics of $X_{1:n}$. To simplify the notation in what follows, we let $X_{(0)} = L$ and $X_{(n+1)} = U$. Then the order statistics of the loss samples in the calibration set partition the interval $[L,U]$ into $n+1$ bins:

$$\begin{cases} I_j = (X_{(j)}, X_{(j+1)}] & \text{for } j = 1,\dots,n, \\ I_0 = [X_{(0)}, X_{(1)}] \end{cases} \tag{3}$$

Since each test loss sample lies in $[L,U]$, it must fall into exactly one of the $n+1$ bins in (3). For clarity of presentation in the main body of the paper, we assume that X contains NoTies, meaning the probability of any two random variables X_i and X_j taking the same value is zero. This assumption is not strictly necessary. Appendix A.1 provides the theoretical guarantees without this restriction, which ensures our algorithm remains applicable even when X contains Ties.

For each bin I_j, we define the occupancy count $N_j = \sum_{t=1}^m \mathbf{1}\{X_t^{\text{test}} \in I_j\}$, $j = 0,1,\dots,n$. By construction, $\sum_{j=0}^n N_j = m$. Hence, the vector $\mathbf{N} = (N_0, N_1, \dots, N_n)$ fully characterizes how the m test samples are distributed across the bins induced by the n observed calibration losses. We now compute the joint distribution of $\mathbf{N}$. A surprising result is that it follows a discrete uniform distribution. This theoretical finding serves as the foundation of our method.

Theorem 3.1. *Under the Assumption 2.1,* $\mathbf{N}$ *follows a discrete uniform distribution, which is*

$$\mathbb{P}\big(N_0 = x_0, N_1 = x_1, ..., N_n = x_n\big) = \frac{m!\cdot n!}{(m+n)!}.$$

The randomness here arises from the sampling of different calibration and test losses.

The proofs for all theoretical results are provided in Appendix A. This is an elegant result, as the joint distribution of $N_0, N_1, \dots, N_n$ is entirely independent of the actual values of these variables. Leveraging this occupancy count, we next construct the UCB $R^+(\lambda)$. Let

$$\mathcal{C}_{m,n} = \left\{ \mathbf{x} = (x_0,\dots,x_n) \in \mathbb{N}_0^{n+1} : \sum_{i=0}^n x_i = m \right\}.$$

Let $S_i(\mathbf{x}) = \sum_{k=0}^i x_k$ denote the cumulative sums, where S_{-1} is defined as 0. By Theorem 3.1, the random vector $\mathbf{N}$ is uniformly distributed over $\mathcal{C}_{m,n}$, and $|\mathcal{C}_{m,n}| = \binom{m+n}{n}$. We further define the upper risk functionals associated with $\mathbf{x}$ by

$$\Phi^+(\mathbf{x};\lambda,X_{1:n}) = \sum_{i=0}^n \int_{\frac{S_{i-1}(\mathbf{x})}{m}}^{\frac{S_i(\mathbf{x})}{m}} \psi(t) X_{(i+1)}\, dt.$$

Intuitively, these upper risk functionals are defined to scale each test loss sample falling into I_j to the maximum value within that bin, and then compute the overall risk across the test samples. This construction stems from the fact that, given the occupancy vector $\mathbf{x}$, the count of test samples falling into each bin becomes known. We can thus scale up the test losses in each bin to the maximum value

within that interval to form an upper risk functional. Utilizing this upper risk functional, we define a CDF of the upper risk:

$$F^+(r;\lambda,X_{1:n}) = \frac{1}{|\mathcal{C}_{m,n}|}\sum_{\mathbf{x}\in\mathcal{C}_{m,n}} \mathbf{1}\{\Phi^+(\mathbf{x};\lambda,X_{1:n}) \le r\}.$$

Thus, defining the UCB as $R^+(\lambda) = (F^+)^{-1}(1-\delta)$, we have

Theorem 3.2. *Given a confidence level $1-\delta \in (0,1)$, $\forall\lambda$, $R^+(\lambda)$ is a valid UCB that satisfies (2).*

However, as m and n increase, calculating the CDF of the upper risk $F^+(r;\lambda,X_{1:n})$ becomes computationally demanding. To address this limitation for large values of m and n, we introduce a Monte-Carlo approach to approximate $F^+(r;\lambda,X_{1:n})$ in the following subsection.

3.2 Monte-Carlo Approach for OQRC

We utilize a Monte Carlo method to mitigate the computational bottlenecks that arise when m and n are excessively large. We first introduce the proposed sampling approach, which exploits a stars-and-bars construction to efficiently sample from $\mathcal{C}_{m,n}$.

Proposition 3.3. *Let $T_1 < \cdots < T_n$ be drawn uniformly from all n-subsets of $\{1,\dots,m+n\}$. Define $T_0 = 0$ and $T_{n+1} = m+n+1$, and set*

$$x_i = T_{i+1} - T_i - 1, \qquad i = 0,1,\dots,n.$$

Then $\mathbf{x} = (x_0,\dots,x_n)$ is uniformly distributed on $\mathcal{C}_{m,n}$.

Based on the sampler yielded by Proposition 3.3, we generate Monte-Carlo replicates

$$\mathbf{x}^{(1)},\dots,\mathbf{x}^{(B)} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{C}_{m,n}),$$

where B is the number of Monte-Carlo samples. For each replicate $\mathbf{x}^{(b)}$, we yield an empirical estimator $\widehat{F}_B^+(r;\lambda,X_{1:n}) = \frac{1}{B}\sum_{b=1}^B \mathbf{1}\{\Phi^+(\mathbf{x}^{(b)};\lambda,X_{1:n}) \le r\}$. The next result demonstrates that the Monte-Carlo approximation converges uniformly to $F^+(r;\lambda,X_{1:n})$.

Theorem 3.4. *For any $\varepsilon > 0$,*

$$\mathbb{P}_{\text{MC}}\left(\sup_{r\in\mathbb{R}}\left|\widehat{F}_B^+(r;\lambda,X_{1:n}) - F^+(r;\lambda,X_{1:n})\right| > \varepsilon\right) \le 2e^{-2B\varepsilon^2}.$$

Theorem 3.4 is a direct consequence of the DKW inequality [20, 21] applied to the i.i.d. Monte-Carlo samples $\{\Phi^+(\mathbf{x}^{(b)};\lambda,X_{1:n})\}_{b=1}^B$. We summarize this procedure to Algorithm 1 in Appendix C.

3.3 Other theoretical results on OQRC

Calibrate the λ. We next select a valid λ to achieve quantile risk control. Adopting the choice of Bates et al. [18], we define $\lambda^* := \inf\{\lambda\in\Lambda : R^+(\lambda) \le \alpha\}$. Then it leads to the following theorem.

Theorem 3.5. *Given a risk level α and a confidence level $1-\delta$, we have*

$$\mathbb{P}\Big(R_{\text{test}}(\lambda^*) \le \alpha\Big) \ge 1-\delta.$$

We formalize this selection process for λ as Algorithm 2 in Appendix C.

Tightness of OQRC. We then show that OQRC achieves tight control by proving that the risk control bounds obtained from OQRC converge to the optimal bounds at a rate of $\mathcal{O}_p(n^{-1/2})$. For any λ, we define the optimal $1-\delta$ bound for risk control as $M^*(\lambda) := \inf\{M : \mathbb{P}(R_{\text{test}}(\lambda) \le M) \ge 1-\delta\}$. Under this optimal bound, selecting $\lambda' = \inf\{\lambda\in\Lambda : M^*(\lambda) \le \alpha\}$ yields a value that coincides with $\lambda'' = \inf\{\lambda\in\Lambda : \mathbb{P}(R_{\text{test}}(\lambda) \le \alpha) \ge 1-\delta\}$, indicating the optimality of M^*. Let f_X denote the probability density function of X. We establish the following asymptotic property.

Theorem 3.6. *For a fixed $\lambda\in\Lambda$ and a fixed m, assume $0 < c \le f_X \le C < \infty$, for $x\in[L,U]$ and some constants $c, C > 0$. Then we have*

$$|R^+(\lambda) - M^*(\lambda)| = \mathcal{O}_p(n^{-1/2}).$$

Table 1: **Performance of quantile risk control methods on 7 datasets across 3 tasks for 2 quantile-based risk measures.** Since CDRC-L fails to achieve valid quantile risk control, the values of RiskGap and AvgSize for this method are not reported. The best results are shown in **bold**.

Dataset	Risk Measure	Methods	Metrics		
			Cov	RiskGap (↓)	AvgSize (↓)
Polyp	CVaR	OrderStats	1.000	0.366±0.018	47.021±2.637
		One-sided BJ	1.000	0.324±0.034	41.907±3.775
		CDRC-L	0.844	-	-
		OQRC(Ours)	0.935	**0.120±0.062**	**20.061±3.441**
	VaR-Interval	OrderStats	1.000	0.362±0.022	41.299±4.188
		One-sided BJ	1.000	0.332±0.027	34.926±4.162
		Two-sided BJ	1.000	0.331±0.026	34.729±4.364
		OQRC(Ours)	0.928	**0.168±0.074**	**17.216±3.410**
MS COCO	CVaR	OrderStats	1.000	0.339±0.012	7.612±0.234
		One-sided BJ	1.000	0.325±0.024	6.845±0.866
		CDRC-L	0.841	-	-
		OQRC(Ours)	0.959	**0.082±0.050**	**3.345±0.255**
	VaR-Interval	OrderStats	1.000	0.390±0.020	7.024±0.851
		One-sided BJ	1.000	0.310±0.074	5.127±0.982
		Two-sided BJ	1.000	0.309±0.075	5.090±0.935
		OQRC(Ours)	0.904	**0.066±0.041**	**3.115±0.237**
Go Emotions	CVaR	OrderStats	1.000	0.372±0.030	15.712±4.184
		One-sided BJ	1.000	0.336±0.042	11.092±2.586
		CDRC-L	0.844	-	-
		OQRC(Ours)	0.925	**0.139±0.074**	**5.439±0.838**
	VaR-Interval	OrderStats	1.000	0.399±0.001	12.035±2.975
		One-sided BJ	1.000	0.398±0.002	9.397±2.028
		Two-sided BJ	1.000	0.398±0.012	9.316±2.001
		OQRC(Ours)	0.937	**0.254±0.107**	**5.018±0.743**

This theorem demonstrates that as the calibration set size n increases, the risk bounds obtained by OQRC asymptotically converge to the optimal bounds, thereby proving the tightness of our method. We note that our tightness result applies only to continuous losses, which may be a restrictive assumption in practice. A brief discussion of this limitation is provided in Appendix D.

4 Experiments

4.1 Experimental setups

Datasets and setup. We evaluate the performance of OQRC across three tasks: tumor segmentation, image multi-label classification, and text emotion recognition. The tumor segmentation evaluation pools data from several open-source gut polyp segmentation datasets, specifically Kvasir [14], Hyper-Kvasir [15], CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib. This extracted collection comprises 798 data points consisting of images and their corresponding tumor masks. In the multi-label classification setting, we utilize the MS COCO dataset [16], which includes 80 object classes and allows multiple classes to appear within a single instance. We sample 2000 data points from this source. The text emotion recognition evaluation employs the Go Emotions dataset [17], which contains Reddit comments annotated with binary labels for 28 fine-grained emotion categories. Each individual comment can possess multiple positive labels. This subset includes 5427 examples. Across all evaluations, we uniformly configure the size of the calibration set to 200 and the size of the test set to 500. We use the false negative proportion (FNP) as our loss function, with a detailed explanation in Appendix E.1. We repeat each experiment 1000 times to report the average OQRC performance.

Model. We employ PraNet [22] for tumor segmentation, ResNet50 [23] for image multi-label classification, and BERT [24] for text emotion recognition. To validate the robustness of our method across various model architectures, we evaluate four different architectures on the image multi-label

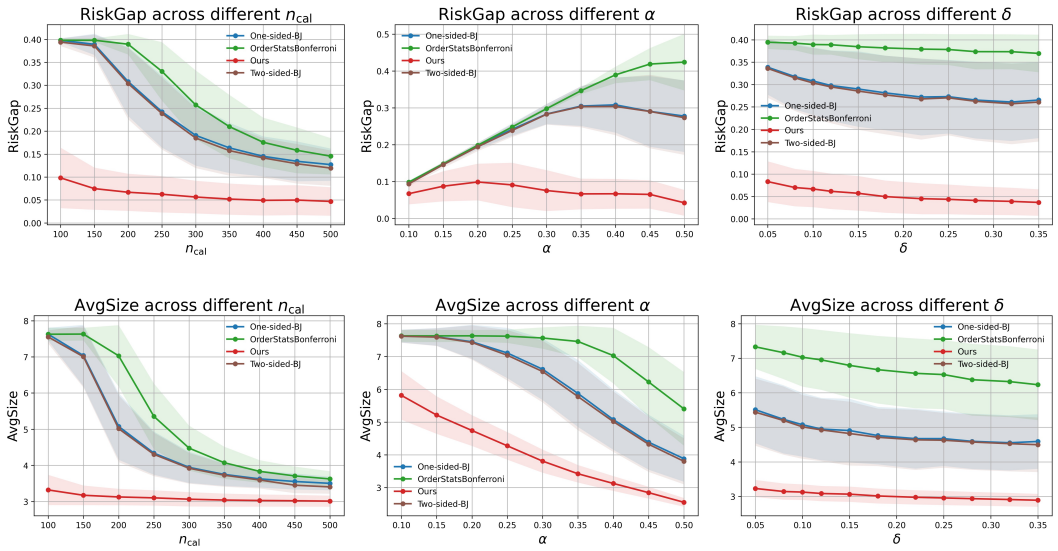


Figure 3: **Performance comparison across various hyperparameter configurations.** The top three figures compare the RiskGap, while the bottom three figures compare the AvgSize. Due to the highly similar performance of One-sided BJ and Two-sided BJ, their curves nearly overlap in the figures.

classification task, which include ResNet [23], EfficientNet [25], TResNet [26], and ConvNeXt [27]. All models are fine-tuned on the corresponding datasets, with training details in Appendix E.2.

Compared methods. We present five methods in total: Order Statistics (**OrderStats**), truncated Berk-Jones (**One-sided BJ**, **Two-sided BJ**), conformal distortion risk control via L-statistics (**CDRC-L**), and our method. The Two-sided BJ is specifically designed for VaR-Interval, while CDRC-L is tailored for CVaR and VaR. Therefore, these two methods are only evaluated in experiments with specific risk measures. Detailed explanations of these methods are provided in Appendix B.

Evaluation metrics. We evaluate the performance of quantile risk control methods using three specific metrics. The first metric is the coverage (**Cov**), defined as the empirical probability that the test risk remains under α across experimental trials. A method is considered valid if the coverage exceeds $1-\delta$, and invalid otherwise. The second metric is the average risk gap (**RiskGap**), defined as the empirical average of the absolute difference between the empirical and target risks. This metric assesses the conservatism of the method. Assuming that the method is valid, a larger RiskGap indicates a more conservative approach, which implies that a smaller gap corresponds to tighter control. Finally, we utilize the average prediction set size (**AvgSize**), defined as the average length of the prediction sets. Similar to the RiskGap, when the method is valid, a smaller AvgSize implies tighter control of the risk. We provide a detailed explanation of these metrics in Appendix E.3.

4.2 Main results

OQRC achieves tighter risk control while maintaining valid coverage. Table 1 compares our proposed method with several baseline methods across coverage, risk gap, and average prediction set size by using the CVaR and the VaR-Interval as risk measures. In the experimental configuration, β is set to 0.8 for the CVaR, while β_{min} and β_{max} for the VaR-Interval are 0.85 and 0.95, respectively. The target risk level α is fixed at 0.4, and the confidence level $1-\delta$ is 0.9. The results demonstrate that our method yields tighter control than OrderStats, One-sided BJ, and Two-sided BJ, which translates to smaller values for the RiskGap and the AvgSize. For instance, on the MS COCO dataset under the VaR-Interval measure, our approach reduces the RiskGap by 78.64% and the AvgSize by 38.80% compared to Two-sided BJ. Meanwhile, CDRC-L fails to satisfy the $1-\delta$ coverage requirement in any of the three tasks, suggesting that the method lacks valid risk control when the size of the calibration set is small. We also observe that the empirical coverage of OQRC modestly exceeds the predefined confidence level 0.9, an experimental result that appears inconsistent with the theoretical

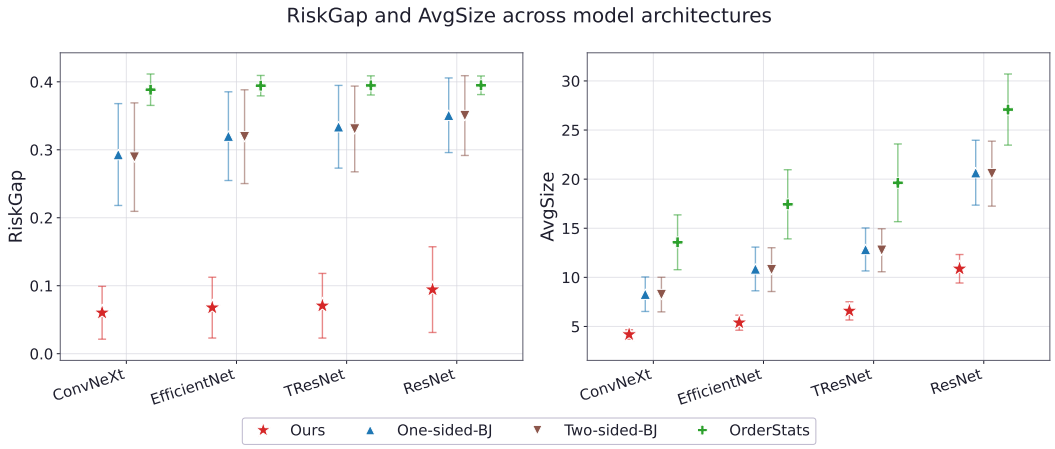


Figure 4: Performance comparison across different model architectures.

intuition presented in Subsection 3.3. Our detailed analysis in Appendix D demonstrates that these empirical findings remain consistent with our theoretical results.

OQRC maintains its effectiveness under various hyperparameter configurations. In Figure 3, we compare the quantile risk control performance across different sizes of the calibration set, target risk levels α, and confidence levels $1-\delta$. We conduct experiments on the MS COCO dataset under the [0.85, 0.95]-VaR-Interval risk measure. The results show that OQRC achieves tighter risk control than the other baseline methods. For instance, compared to alternative approaches, our approach reduces the RiskGap by 82.54% and the AvgSize by 44.66% on average across varying values of δ. Appendix H.3 provides additional experiments that involve other risk measures and datasets.

OQRC performs effectively across different model architectures. Figure 4 compares the performance of our approach against baseline methods under various model architectures. We conduct experiments on the MS COCO dataset under the [0.85, 0.95]-VaR-Interval risk measure. Across all these models, OQRC decreases the average RiskGap from 0.347 to 0.075 and reduces the average AvgSize from 15.226 to 6.849, consistently outperforming other baseline methods. These results highlight the model-agnostic nature of our method. Appendix H.4 provides further performance comparisons between our approach and the baseline methods by using alternative risk measures.

5 Discussion

Unconditional quantile risk control. Achieving unconditional quantile risk control, which requires ensuring that $R_\psi(F_X) \leq \alpha$, is a challenging problem [6]. Yeh et al. [28] propose a training-based solution to achieve unconditional control over the optimized certainty equivalent (OCE) risks. However, training-free methods that achieve unconditional control over QBRMs are still lacking. Our method offers a solution to this problem, leading to the following theorem.

Theorem 5.1. *For any fixed $\lambda \in \Lambda$, let $\mathbf{x}$ denote the occupancy vector induced by the calibration losses $X_{1:n}(\lambda)$ and the test losses $X^{\text{test}}_{1:m}(\lambda)$. Then*

$$\mathbb{E}_{\text{cal,test}}\left[R_{\text{test}}(\lambda)\right] \leq \mathbb{E}_{\text{cal,test}}\left[\Phi^+\left(\mathbf{x};\lambda,X_{1:n}\right)\right]. \tag{4}$$

The expectation on the right-hand side is taken with respect to the actual joint distribution of the occupancy vector and the calibration losses.

Consequently, we select λ as $\lambda^* := \inf\{\lambda \in \Lambda : \mathbb{E}\left[\Phi^+\left(\mathbf{x};\lambda,X_{1:n}\right)\right] \leq \alpha\}$, then we achieve the unconditional quantile risk $\mathbb{E}[R_{\text{test}}(\lambda^*)] \leq \alpha$. The procedure for selecting λ is detailed in Subsection 3.3. As demonstrated in Table 2, our method successfully bounds $\mathbb{E}\left[R_{\text{test}}\right]$ strictly below the threshold α across three distinct tasks under various predefined values of α.

Relation to conformal risk control. We next demonstrate that the conformal risk control method [6] is a special case of Theorem 5.1. By setting $m = 1$ and selecting the expected loss as the risk

	Dataset	Risk Level	Empirical Risk		Dataset	Risk Level	Empirical Risk
CVaR	Polyp	$\alpha = 0.4$	$0.379(\pm 0.021)$	Interval	Polyp	$\alpha = 0.4$	$0.382(\pm 0.018)$
		$\alpha = 0.2$	$0.181(\pm 0.019)$			$\alpha = 0.2$	$0.185(\pm 0.015)$
		$\alpha = 0.1$	$0.077(\pm 0.023)$			$\alpha = 0.1$	$0.091(\pm 0.009)$
	MS COCO	$\alpha = 0.4$	$0.385(\pm 0.015)$		MS COCO	$\alpha = 0.4$	$0.393(\pm 0.007)$
		$\alpha = 0.2$	$0.178(\pm 0.022)$			$\alpha = 0.2$	$0.189(\pm 0.011)$
		$\alpha = 0.1$	$0.081(\pm 0.019)$			$\alpha = 0.1$	$0.091(\pm 0.009)$
	Go Emotions	$\alpha = 0.4$	$0.387(\pm 0.013)$		Go Emotions	$\alpha = 0.4$	$0.369(\pm 0.031)$
		$\alpha = 0.2$	$0.186(\pm 0.014)$			$\alpha = 0.2$	$0.179(\pm 0.021)$
		$\alpha = 0.1$	$0.088(\pm 0.012)$			$\alpha = 0.1$	$0.078(\pm 0.022)$

Table 2: **Unconditional quantile risk control performance via OQRC.** The values in $(\cdot)$ denote the difference between Risk Level and Empirical Risk, where a positive value indicates valid control.

measure, the space $\mathcal{C}$ is formed by $(n+1)$-dimensional one-hot vectors. Let $\mathbf{x}$ be a one-hot vector where the r-th component is 1 and all other components are 0. Under this formulation, we obtain $\Phi^+(\mathbf{x}; \lambda, X_{1:n}) = X_{(r+1)}$. It follows that $\mathbb{E}\big[\Phi^+(\mathbf{x}; \lambda, X_{1:n})\big] = \sum_{i=0}^{n} \frac{1}{n+1} X_{(i+1)}$. Given that $R_{\text{test}} = X^{\text{test}}$, we substitute these relations into Theorem 5.1 and choose $\lambda^* := \inf\big\{\lambda \in \Lambda : \sum_{i=0}^{n} \frac{1}{n+1} X_{(i+1)}(\lambda) \leq \alpha\big\}$. Then we recover the CRC procedure:

$$\mathbb{E}\big[X^{\text{test}}(\lambda^*)\big] \leq \alpha.$$

The limiting case as $m \to \infty$. We further discuss the OQRC method when $m \to \infty$. This condition arises when controlling the risk over the entire loss distribution. As $m \to \infty$, we obtain unconditional quantile risk control $R_\psi(F_X; \lambda) \leq \alpha$. A detailed analysis is presented in Appendix F.

Distribution shift. Our finite-sample guarantees rely on the assumption that the calibration and test losses are i.i.d. In the presence of distribution shift, the reweighting technique developed for traditional expected risk measures [29, 30, 31] could potentially be adapted. We further evaluate OQRC under covariate shift and class-frequency shift to examine how its performance degrades when the i.i.d. condition is violated. The results are reported in Appendix G.

How to choose the test size m. Since our method depends on the choice of the test size m, we conduct experiments with different values of m in Appendix H.1 to provide practical guidance on how to select m. We recommend that if practitioners know the test size in advance, they should explicitly set to this prior value. Otherwise, practitioners ought to estimate the test size or adopt the formulation provided in Appendix F to address scenarios where the test size is infinitely large.

Sensitivity to the number of Monte Carlo occupancy samples. In Appendix H.2, we discuss the sensitivity to the number of Monte Carlo occupancy samples. From a practical perspective, we recommend setting the number of Monte Carlo samples to at least 10,000.

6 Related Work

Distribution-free uncertainty quantification. This line of research originates from conformal prediction [32, 33, 34], which generates prediction sets that contain ground-truth labels with a desired coverage guarantee. Subsequent research [6, 35, 36] extends conformal prediction to accommodate more general loss functions. These methods formalize the expected loss associated with the output of the model as the risk, which quantifies the potential unreliability of the predictions. Such a risk control framework is widely applied in domains including large language models [37, 38, 39], computer vision [40, 41, 42], and medicine [43, 44, 45]. In this work, we propose a novel method for distribution-free uncertainty quantification to control a general notion of risk beyond expected loss.

Quantile risk control. In some real-world scenarios, it is necessary to control the tail of the loss distribution beyond the expected loss [46, 47, 48, 49, 50, 51]. To control this general notion of risk, Snell et al. [9] propose a quantile risk control framework. Their method employs a one-sided test

statistic to control the quantile risk, but such a procedure tends to overestimate the prediction risk and thus produce conservative results. Chen et al. [12] achieve tighter risk control via L-statistics, yet their guarantees rely on asymptotic properties as the calibration set size grows to infinity and do not provide finite-sample guarantees. To the best of our knowledge, we are the first to propose a quantile risk control method that provides both finite-sample guarantees and theoretically tight control.

Other works on risk control. Some works control the risk through model training methods [28, 52, 53, 54, 55]. These approaches formulate risk as a learning objective, training the model to either minimize the prediction risk or constrain it below a user-specified threshold. OQRC is a post-hoc method without model training, which is computationally more efficient than these approaches.

7 Conclusion

In this work, we propose Occupancy-based Quantile Risk Control (OQRC), a novel training-free method that achieves finite-sample quantile risk control. The key idea of OQRC is to formulate risk control as a finite-occupancy problem by partitioning the loss space using ordered calibration losses. Theoretically, we establish the tightness of OQRC by proving that the risk control bounds yielded by OQRC converge to the optimal bounds. Extensive empirical evaluations validate the effectiveness of OQRC. Our approach also resolves the problem posed by Angelopoulos et al. [6], enabling users to control the prediction risk without configuring a δ. We hope that our approach provides valuable insights and methodologies for controlling quantile risk across a broader range of scenarios.

Limitation. Our method applies only to monotone loss functions, whereas practical applications may encounter general loss functions. We leave the investigation of this setting to future work.

8 Acknowledgements

This research is supported by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515011367). This project is also supported by the Jiangsu Provincial Key Discipline Construction Project (Statistics) and open project of Joint Lab for Statistics and Finance (Grant No. 2025JLSF101). We thank Shenzhen Guangming Science City Development and Construction Co., Ltd. and its Material Genome Big-Science Facilities Platform, for providing technical support and assistance in data collection and analysis.

References

[1] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[2] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[4] Quentin Waymel, Sammy Badr, Xavier Demondion, Anne Cotten, and Thibaut Jacques. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagnostic and interventional imaging*, 100(6):327–336, 2019.

[5] Pouyan Esmaeilzadeh, Tala Mirzaei, and Spurthy Dharanikota. Patients' perceptions toward human–artificial intelligence interaction in health care: experimental study. *Journal of medical Internet research*, 23(11):e25856, 2021.

[6] Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, volume 2024, pages 55198–55218, 2024.

[7] Kerr Ding, Jiaqi Luo, and Yunan Luo. Leveraging conformal prediction to annotate enzyme function space with limited false positives. *PLOS Computational Biology*, 20(5):e1012135, 2024.

[8] Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Conformal linguistic calibration: Trading-off between factuality and specificity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] Jake Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. In *The Eleventh International Conference on Learning Representations*, 2023.

[10] Jacques Longerstaey and Martin Spencer. Riskmetricstm—technical document. *Morgan Guaranty Trust Company of New York: New York*, 51:54, 1996.

[11] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[12] Catherine Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. Conformal tail risk control for large language model alignment. In *Forty-second International Conference on Machine Learning*, 2025.

[13] Frederick Mosteller. On some useful "inefficient" statistics. In *Selected Papers of Frederick Mosteller*, pages 69–100. Springer, 2006.

[14] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.

[15] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[17] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054, 2020.

[18] Stephen Bates, Anastasios Nikolas Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68:43:1–43:34, 2021.

[19] William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 2. John Wiley & Sons, 1991.

[20] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[21] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[22] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[26] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1400–1409, 2021.

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[28] Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[30] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

[31] Emmanuel Candes, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

[32] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer, 2002.

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

[34] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

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

[36] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

[37] Metod Jazbec, Alexander Timans, Tin H Veljković, Kaspar Sakmann, Dan Zhang, Christian A Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. *Advances in Neural Information Processing Systems*, 37:129825–129854, 2024.

[38] Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Sconu: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19052–19075, 2025.

[39] Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. Coin: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33764–33772, 2026.

[40] Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3321–3333, 2023.

[41] Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, pages 33940–33960. PMLR, 2023.

[42] Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3574–3584, 2024.

[43] Mariya Lysenkova Wiklander, Dave Zachariah, Olga Krali, and Jessica Nordlund. Error reduction in leukemia machine learning classification with conformal prediction. *JCO Clinical Cancer Informatics*, 9:e2400324, 2025.

[44] Mengxia Dai, Wenqian Luo, and Tianyang Li. Statistical management of the false discovery rate in medical instance segmentation based on conformal risk control. *arXiv preprint arXiv:2504.04482*, 2025.

[45] Ron S Boger, Seyone Chithrananda, Anastasios N Angelopoulos, Peter H Yoon, Michael I Jordan, and Jennifer A Doudna. Functional protein mining with conformal guarantees. *Nature Communications*, 16(1):85, 2025.

[46] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[47] Kevin Dowd. Using order statistics to estimate confidence intervals for quantile-based risk measures. *Journal of Derivatives*, 17(3):9, 2010.

[48] Ahmad Ghasemi, Houman Jamshidi Monfared, Abdolah Loni, and Mousa Marzband. Cvar-based retail electricity pricing in day-ahead scheduling of microgrids. *Energy*, 227:120529, 2021.

[49] Amirmohammad Farzaneh, Sangwoo Park, and Osvaldo Simeone. Quantile learn-then-test: Quantile-based risk control for hyperparameter optimization. *IEEE Signal Processing Letters*, 31:3044–3048, 2024.

[50] Zhengyu Wang, Yingtong Liu, Fang Chen, Yanyi Fu, Ming Yang, Zhuhua Chen, and Jie Jiang. Voltage risk assessment of wind power grid connection systems based on cvar. In *2025 5th International Conference on New Energy and Power Engineering (ICNEPE)*, pages 574–581. IEEE, 2025.

[51] Tenghan Zhong. Proxy-reliance control in conformal recalibration of one-sided value-at-risk. *arXiv preprint arXiv:2603.22569*, 2026.

[52] Jaeho Lee, Sejun Park, and Jinwoo Shin. Learning bounds for risk-sensitive learning. *Advances in Neural Information Processing Systems*, 33:13867–13879, 2020.

[53] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[54] Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. Supervised learning with general risk functionals. In *International Conference on Machine Learning*, pages 12570–12592. PMLR, 2022.

[55] Kevin He, David Adam, Sarah Han-Oh, and Anqi Liu. Training-aware risk control for intensity modulated radiation therapies quality assurance with conformal prediction. In Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang, editors, *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 456–470. PMLR, 15–16 Dec 2025.

[56] Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pages 1504–1512. PMLR, 2024.

[57] C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 722–726, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[58] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

[59] Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

[60] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[61] Amit Moscovich, Boaz Nadler, and Clifford Spiegelman. On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329 – 2354, 2016.

A Proofs

A.1 Proof of Theorem 3.1

Proof. Our proof draws inspiration from the work of Gazin et al. [56] The theoretical novelty of this proof lies in applying the methodology of conformal p-values [57, 58, 59] to the risk control framework and relaxing the previous 'NoTies' assumption, thereby making it more applicable to real-world scenarios. The first two steps are proven under the assumption that X has 'NoTies', while the third step relaxes this 'NoTies' assumption.

Step 1: Prove that $\mathbb{P}(\cap_{i=1}^{m}\{X_i^{\text{test}} \in I_{j_i}\}) = \frac{n!\Pi_{k=0}^{n}N_k!}{(m+n)!}$, where $N_k = \sum_{i=1}^{m}\mathbf{1}\{j_i = k\}$.

Let $X_{1:n+m} = X_{1:n} \cup X_{1:m}^{\text{test}}$, where the first n values are loss samples from the calibration set, and the remaining m values are loss samples from the test set. Let σ be a permutation such that

$$X_{\sigma(1)} < X_{\sigma(2)} < \cdots < X_{\sigma(n+m)}.$$

Thus, on the event $\cap_{i=1}^{m}\{X_i^{\text{test}} \in I_{j_i}\}$, the permuted sequence $X_{1:n+m}$ under σ can be uniquely represented as a partition of the following form:

$$\begin{aligned}
&\underbrace{X_{\sigma(1)} < \cdots < X_{\sigma(N_0)}}_{N_0 \text{ test loss}} < \underbrace{X_{\sigma(N_0+1)}}_{\text{calibration loss}} < \\
&\underbrace{X_{\sigma(N_0+2)} < \cdots < X_{\sigma(N_0+N_1+1)}}_{N_1 \text{ test loss}} < \underbrace{X_{\sigma(N_0+N_1+2)}}_{\text{calibration loss}} < \cdots < \\
&\underbrace{X_{\sigma(\sum_{i=0}^{n-1}N_i+n)} < \cdots < X_{\sigma(\sum_{i=0}^{n}N_i+n)}}_{N_n \text{ test loss}} < U.
\end{aligned}$$

Since the permutations of $\{1, 2, \ldots, n+m\}$ are uniformly distributed, the probability of the event $\cap_{i=1}^{m}\{X_i^{\text{test}} \in I_{j_i}\}$ follows the classical probability model, yielding:

$$\mathbb{P}(\cap_{i=1}^{m}\{X_i^{\text{test}} \in I_{j_i}\}) = \frac{n!N_0!\cdot N_1!\cdot \ldots \cdot N_n!}{(m+n)!} = \frac{n!\Pi_{k=0}^{n}N_k!}{(m+n)!}.$$

Step2: Prove that $\mathbb{P}\big(N_0 = x_0, N_1 = x_1, .., N_n = x_n\big) = \frac{m!\cdot n!}{(m+n)!}$.

$$\begin{aligned}
\mathbb{P}\big(N_0 = x_0, N_1 = x_1, .., N_n = x_n\big) &= \sum_{\forall k,\ (\sum_{i=1}^{m}\mathbf{1}\{j_i=k\})=x_k} \mathbb{P}(\cap_{i=1}^{m}\{X_i^{\text{test}} \in I_{j_i}\}) \\
&= \sum_{\forall k,\ (\sum_{i=1}^{m}\mathbf{1}\{j_i=k\})=x_k} \frac{n!\Pi_{k=0}^{n}x_k!}{(m+n)!} \\
&= \frac{m!}{\Pi_{k=0}^{n}x_k!}\cdot\frac{n!\Pi_{k=0}^{n}x_k!}{(m+n)!} \\
&= \frac{m!\cdot n!}{(m+n)!}.
\end{aligned}$$

The third equality holds because we assign a label from 0 to n to each test loss. The label 0 appears x_0 times, ..., and the label n appears x_n times. Since there are m test losses in total, there are $m!$ ways to arrange these labels. However, because swapping identical labels does not produce a new sequence, we must divide by $\prod_{k=0}^{n}x_k!$.

Step 3: Relax the 'NoTies' Assumption on X.

Let $S_i := S(Z_i) = (\ell(h(Z_i), Y_i), Z_i) \in \mathbb{R}^{d+1}$, and $X_i = \ell(h(Z_i), Y_i)$. We define an order relation on the outputs of S such that $S_1 < S_2$ if and only if there exists an integer $q \in [0, d+1)$ where the first q components of S_1 and S_2 are equal, and the $(q+1)$-th component of S_1 is strictly less than that of S_2. Furthermore, $S_1 = S_2$ is defined as all components being identical. It is easy to verify

that this order relation $<$ is a strict total order. It is straightforward to show that the S_i satisfy the 'NoTies' property, even if the X_i may have Ties. Following this, we define $n+1$ intervals as:

$$\begin{cases} I_j = (S_{(j)}, S_{(j+1)}] & \text{for } j = 1, \dots, n, \\ I_0 = [S_{(0)}, S_{(1)}] \end{cases} \tag{5}$$

Here, $S \in (S_{(j)}, S_{(j+1)}]$ is equivalent to $S_{(j)} < S$ and $S \leq S_{(j+1)}$. Note the property that $S_1 < S_2$ implies $X_1 \leq X_2$. Substituting (5) into the previous conclusion with some minor modifications (replacing $S_{(i)}$ with $X_{(i)}$ when calculating R^+), we obtain the OQRC method under the setting where X contains ties. Thus, we completes the proof of Theorem 3.1. □

A.2 Proof of Theorem 3.2

Proof. Fix an arbitrary $\lambda \in \Lambda$. We suppress the dependence on λ in the notation. Let $K := |\mathcal{C}_{m,n}| = \binom{m+n}{n}$. Let $\Phi^{+,i} = \Phi^+(\mathbf{N}^i; X_{1:n})$ denote the upper risk for the i-th occupancy vector, and let $\Phi^{+,(i)}$ represent the corresponding order statistics. We assume without loss of generality that there are no ties among $\Phi^{+,1}, \Phi^{+,2}, \dots, \Phi^{+,K}$. If ties exist, they can be broken using the auxiliary pair $(\Phi^{+,i}, U_i)$, where $U_i \sim \mathrm{Unif}(0,1)$ is drawn independently of $\Phi^{+,i}$. Let $\mathrm{Rank}(\Phi^{+,i})$ be the rank of $\Phi^{+,i}$ in ascending order. Theorem 3.1 indicates that $\mathrm{Rank}(\Phi^{+,i}) \sim \mathrm{Unif}\{1, 2, \dots, K\}$. For the occupancy vector $\mathbf{N}$ associated with each R_{test}, the inequality $R_{\text{test}} \leq \Phi^+(\mathbf{N}; X_{1:n})$ holds. Let $k := \lceil (1-\delta)K \rceil$. This implies the inclusion relation

$$\{\Phi^+(\mathbf{N}; X_{1:n}) \leq \Phi^{+,(k)}\} \subseteq \{R_{\text{test}} \leq \Phi^{+,(k)}\}.$$

Recall that $R^+ = (F^+)^{-1}(1-\delta)$. Then $R^+ \geq \Phi^{+,(k)}$. We thus obtain

$$\mathbb{P}\big(R_{\text{test}} \leq R^+\big) \geq \mathbb{P}\big(\Phi^+(\mathbf{N}; X_{1:n}) \leq R^+\big) \geq 1 - \delta.$$

Hence, R^+ serves as a valid $(1-\delta)$ upper confidence bound for R_{test}. □

A.3 Proof of Proposition 3.3

Proof. Let

$$\mathcal{T}_{m,n} := \Big\{(t_1, \dots, t_n) \in \{1, \dots, m+n\}^n : \ 1 \leq t_1 < \cdots < t_n \leq m+n\Big\}.$$

By assumption, $(T_1, \dots, T_n)$ is uniformly distributed on $\mathcal{T}_{m,n}$. Define the map

$$\Gamma : \mathcal{T}_{m,n} \to \mathcal{C}_{m,n}, \qquad \Gamma(t_1, \dots, t_n) = (x_0, \dots, x_n),$$

where

$$x_i := t_{i+1} - t_i - 1, \qquad i = 0, \dots, n,$$

with the conventions $t_0 := 0$ and $t_{n+1} := m+n+1$. This is exactly the stars-and-bars map.

The inverse map is explicit. Given $\mathbf{x} = (x_0, \dots, x_n) \in \mathcal{C}_{m,n}$, define

$$t_i := i + \sum_{k=0}^{i-1} x_k, \qquad i = 1, \dots, n.$$

Then $(t_1, \dots, t_n) \in \mathcal{T}_{m,n}$ and $\Gamma(t_1, \dots, t_n) = \mathbf{x}$. Hence Γ is a bijection.

Since $(T_1, \dots, T_n)$ is uniform on $\mathcal{T}_{m,n}$, its image

$$x_i = T_{i+1} - T_i - 1, \qquad i = 0, \dots, n,$$

is uniform on $\mathcal{C}_{m,n}$. That is,

$$\mathbf{x} = (x_0, \dots, x_n) \sim \mathrm{Unif}(\mathcal{C}_{m,n}).$$

□

A.4 Proof of Theorem 3.5

Proof. The proof of this theorem is identical to that of Theorem A.1 in Bates et al. [18]. □

A.5 Proof of Theorem 3.6

Proof. We redefine the optimal bound as $M^* = \inf\{M : \mathbb{P}\big(R_\psi(\frac{1}{m}\sum_{j=1}^m \delta_{Y_j}) \le M\big) \ge 1-\delta\}$. This formulation is equivalent to the definition of M^* in Subsection 3.3, but it is more convenient for our subsequent proof. We divide our proof into four steps.

Step 1: Lipschitz property of the quantile risk.

Let $a = (a_1, \dots, a_m)$ and $b = (b_1, \dots, b_m)$ be two vectors in $[L, U]^m$. Let $a_{(1)} \le \cdots \le a_{(m)}, b_{(1)} \le \cdots \le b_{(m)}$ be their order statistics. Define the empirical quantile functions $Q_a(t) = a_{(j)}$ for $t \in \big(\frac{j-1}{m}, \frac{j}{m}\big]$, and similarly for Q_b. If $\max_{1\le j\le m} |a_{(j)} - b_{(j)}| \le \varepsilon$, then $\sup_{t\in[0,1]} |Q_a(t) - Q_b(t)| \le \varepsilon$. Since ψ is a probability measure,

$$\left|\int_0^1 Q_a(t)\,\psi(dt) - \int_0^1 Q_b(t)\,\psi(dt)\right| \le \int_0^1 |Q_a(t) - Q_b(t)|\,\psi(dt) \le \varepsilon.$$

Therefore, the quantile risk is 1-Lipschitz with respect to the supremum norm of the sorted losses.

Step 2: Uniform convergence of the calibration quantile function.

Let $\widehat{F}_n$ be the empirical CDF of $X_1, \dots, X_n$. By the Dvoretzky–Kiefer–Wolfowitz inequality,

$$\sup_{x\in[L,U]} |\widehat{F}_n(x) - F(x)| = \mathcal{O}_p(n^{-1/2}).$$

Because $f(x) \ge c > 0$ on $[L, U]$, the population quantile function $Q = F^{-1}$ is $1/c$-Lipschitz:

$$|Q(u) - Q(v)| \le \frac{1}{c}|u - v|, \qquad u, v \in [0, 1].$$

The standard inverse-CDF perturbation bound gives

$$\sup_{p\in[0,1]} \left|\widehat{Q}_n^+(p) - Q(p)\right| \le \frac{1}{c}\left(\sup_{x\in[L,U]} |\widehat{F}_n(x) - F(x)| + \frac{1}{n}\right).$$

where $\widehat{Q}_n^+(p) = X_{(\min\{n+1, \lceil (n+1)p\rceil\})}$. Hence

$$D_n := \sup_{p\in[0,1]} \left|\widehat{Q}_n^+(p) - Q(p)\right| = \mathcal{O}_p(n^{-1/2}).$$

Step 3: Coupling the occupancy grid with the test order statistics.

For $\mathbf{x} \in \mathcal{C}_{m,n}$, form the nondecreasing index sequence $0 \le J_1(\mathbf{x}) \le \cdots \le J_m(\mathbf{x}) \le n$ by listing the index i exactly x_i times. Thus, x_i is the number of indices j for which $J_j(\mathbf{x}) = i$. As $\mathbf{x}$ is uniformly distributed on $C_{m,n}$, then $V_j = \frac{J_j(x)+1}{n+1}$ is uniformly distributed on the ordered grid

$$\mathcal{G}_{m,n} = \left\{(v_1, \dots, v_m) : v_j = \frac{k_j + 1}{n+1},\ 0 \le k_1 \le \cdots \le k_m \le n\right\}.$$

Let $0 \le U_{(1)} \le \cdots \le U_{(m)} \le 1$ be the order statistics of m i.i.d. $\mathrm{Unif}(0, 1)$ random variables. The ordered-grid distribution on $\mathcal{G}_{m,n}$ is a regular $\mathcal{O}(n^{-1})$ discretization of the uniform distribution on the simplex $\Delta_m = \{0 \le u_1 \le \cdots \le u_m \le 1\}$. Therefore, for fixed m, there exists a coupling of $(V_1, \dots, V_m)$ and $(U_{(1)}, \dots, U_{(m)})$ such that

$$\max_{1\le j\le m} |V_j - U_{(j)}| \le \frac{A_m}{n}$$

for a constant $A_m < \infty$ depending only on m. Under this coupling, since Q is $1/c$-Lipschitz, then

$$\max_{1\le j\le m} |Q(V_j) - Q(U_{(j)})| \le \frac{A_m}{c\,n}.$$

Step 4: Comparing the exact OQRC risk with the optimal test risk.

For $x \in C_{m,n}$, the vector used by the exact occupancy upper bound is $(X_{(J_1(x)+1)}, \ldots, X_{(J_m(x)+1)})$. Equivalently, $X_{(J_j(x)+1)} = \widehat{Q}_n^+(V_j)$. Therefore,

$$\max_{1\le j\le m} \left|X_{(J_j(x)+1)} - Q(U_{(j)})\right| \le \max_{1\le j\le m} \left|\widehat{Q}_n^+(V_j) - Q(V_j)\right| + \max_{1\le j\le m} \left|Q(V_j) - Q(U_{(j)})\right|.$$

Using the bounds from Steps 2 and 3,

$$\max_{1\le j\le m} \left|X_{(J_j(x)+1)} - Q(U_{(j)})\right| \le D_n + \frac{A_m}{c\,n}.$$

By the Lipschitz property from Step 1,

$$\left|\Phi_n^+(x) - R_\psi\left(\frac{1}{m}\sum_{j=1}^m \delta_{Q(U_{(j)})}\right)\right| \le D_n + \frac{A_m}{c\,n}.$$

Since $Q(U_{(1)}), \ldots, Q(U_{(m)})$ are the order statistics of m i.i.d. samples from X, the random variable $R_\psi(\frac{1}{m}\sum_{j=1}^m \delta_{Q(U_{(j)})})$ has the same distribution as $R_\psi(\frac{1}{m}\sum_{j=1}^m \delta_{Y_j})$. Hence its $(1-\delta)$ quantile is exactly M^*. Then, if two random variables A and B can be coupled so that $|A - B| \le \varepsilon$ almost surely, their $(1-\delta)$ quantiles differ by at most ε. Applying this fact with $A = \Phi_n^+(x)$, $B = R_\psi(\frac{1}{m}\sum_{j=1}^m \delta_{Y_j})$, $\varepsilon = D_n + \frac{A_m}{c\,n}$, we obtain

$$\left|R^+ - M^*\right| \le D_n + \frac{A_m}{c\,n}.$$

Since $D_n = \mathcal{O}_p(n^{-1/2})$, it follows that

$$\left|R^+ - M^*\right| = \mathcal{O}_p(n^{-1/2}).$$

This proves the theorem. □

A.6 Proof of Theorem 5.1

Proof. Fix $\lambda \in \Lambda$ and suppress the dependence of the losses on λ throughout the proof. Let $X_{(1)} \le \cdots \le X_{(n)}$ be the calibration order statistics, and set $X_{(n+1)} = U$. Define the bins

$$I_0 = [L, X_{(1)}], \qquad I_i = (X_{(i)}, X_{(i+1)}], \quad i = 1, \ldots, n.$$

These bins form a partition of $[L, U]$, with some bins possibly empty.

Write $T_j = X_j^{\text{test}}$, $j = 1, \ldots, m$, and let $T_{(1)} \le \cdots \le T_{(m)}$ be the test order statistics. For the empirical test distribution $\widehat{F}_m(t) = \frac{1}{m}\sum_{j=1}^m \mathbf{1}\{T_j \le t\}$, its generalized inverse satisfies

$$\widehat{F}_m^{-1}(p) = T_{(r)} \qquad \text{for } p \in \left(\frac{r-1}{m}, \frac{r}{m}\right], \quad r = 1, \ldots, m.$$

Define $w_r = \int_{(r-1)/m}^{r/m} \psi(p)\,dp, \qquad r = 1, \ldots, m$. Since ψ is nonnegative and integrates to one, $w_r \ge 0$ and $\sum_{r=1}^m w_r = 1$. Consequently,

$$R_{\text{test}}(\lambda) = \int_0^1 \psi(p)\widehat{F}_m^{-1}(p)\,dp = \sum_{r=1}^m w_r T_{(r)}. \tag{6}$$

Let

$$x_i = \sum_{j=1}^m \mathbf{1}\{T_j \in I_i\}, \qquad S_i = \sum_{k=0}^i x_k, \qquad S_{-1} = 0.$$

For each realization of the calibration and test samples, S_i is the number of test losses contained in $I_0 \cup \cdots \cup I_i$. Because the bins are ordered and disjoint, we have

$$S_{i-1} < r \le S_i \quad \Longrightarrow \quad T_{(r)} \in I_i \quad \Longrightarrow \quad T_{(r)} \le X_{(i+1)}.$$

Using this observation and the nonnegativity of the weights, we obtain the following inequality:

$$\begin{aligned} R_{\text{test}}(\lambda) &= \sum_{i=0}^{n} \sum_{r=S_{i-1}+1}^{S_i} w_r T_{(r)} \\ &\le \sum_{i=0}^{n} X_{(i+1)} \sum_{r=S_{i-1}+1}^{S_i} w_r \\ &= \sum_{i=0}^{n} X_{(i+1)} \int_{S_{i-1}/m}^{S_i/m} \psi(p)\, dp \\ &= \Phi^+(\mathbf{x}; \lambda, X_{1:n}). \end{aligned} \tag{7}$$

Here, a sum over an empty index set is interpreted as zero. In particular, the argument remains valid when some losses coincide.

The order statistics, occupancy counts, and quantities in (7) are measurable functions of the samples. Moreover, both sides of (7) are nonnegative weighted averages of values in $[L, U]$, with weights summing to one. Thus,

$$L \le R_{\text{test}}(\lambda) \le U, \qquad L \le \Phi^+(\mathbf{x}; \lambda, X_{1:n}) \le U,$$

so both random variables are integrable. Taking expectations in (7) with respect to the joint distribution of the calibration and test samples proves (4). □

B Other methods

B.1 Methods in Quantile Risk Control

We first introduce the general framework of quantile risk control [9]. Given a calibration set $X_{1:n} = \{X_1, X_2, \ldots, X_n\}$, we denote its ascending order statistics as $X_{(1)} \le \ldots \le X_{(n)}$. Let $U_1, U_2, \ldots, U_n \overset{iid}{\sim} \text{Unif}(0,1)$, where $U_{(i)}$ denotes their corresponding order statistics. The procedure begins by selecting a one-sided test statistic of the form $S \triangleq \min_{1 \le i \le n} s_i(U_{(i)})$. Next, we compute the critical value $s_\delta = \inf\{r : \mathbb{P}(S \ge r) \ge 1 - \delta\}$. We then construct an empirical bounding function $\hat{F}_n$ defined as

$$\hat{F}n(x) \triangleq \begin{cases} 0, & \text{for } x < X(1) \\ s_1^{-1}(s_\delta), & \text{for } X_{(1)} \le x < X_{(2)} \\ \vdots & \\ s_n^{-1}(s_\delta), & \text{for } X_{(n)} \le x < x^+ \\ 1, & \text{for } x^+ \le x, \end{cases}$$

where $x^+ \in \mathbb{R} \cup \{\infty\}$ is an upper bound such that $F(x^+) = 1$. This construction guarantees the probabilistic bound

$$\mathbb{P}(R_\psi(F) \le R_\psi(\hat{F}_n)) \ge 1 - \delta.$$

Within this framework, different choices of the one-sided test statistic yield various risk control methods. The one-sided Kolmogorov-Smirnov (**KS**) statistic [60] sets $s_i(u) = u - \frac{i}{n}$. The Berk-Jones (**BJ**) statistic [61] employs $s_i(u) = I_u(i, n - i + 1)$, where $I_u(i, n - i + 1)$ denotes the cumulative distribution function of the $\text{Beta}(i, n - i + 1)$ distribution evaluated at u. Furthermore, the order statistics (**OrderStats**) approach defines $s_i(u) = \mathbf{1}\{i = k\}\, I_u(k, n - k + 1) + \mathbf{1}\{i \ne k\}$. Extensions include the one-sided truncated Berk-Jones (**One-sided BJ**) statistic, which corresponds to $s_i(u) = \mathbf{1}\{i \ge k\}\, I_u(i, n-i+1) + \mathbf{1}\{i < k\}$, and the two-sided truncated Berk-Jones (**Two-sided BJ**) statistic, which specifies $s_i(u) = \mathbf{1}\{k \le i \le l\}\, I_u(i, n - i + 1) + \mathbf{1}\{i < k \text{ or } l < i\}$.

The selection of k and l in the OrderStats, One-sided BJ, and Two-sided BJ methods depends on the β parameter in β-CVaR and the $\beta_{\min}$ and $\beta_{\max}$ parameters in the VaR-Interval measure. We define the statistics $M_{n,k}^+ \triangleq \min_{k \le i \le n} I_{U_{(i)}}(i, n-i+1)$ and $M_{n,k,l}^+ \triangleq \min_{k \le i \le l} I_{U_{(i)}}(i, n-i+1)$. Denoting s_δ^k as the critical value of $M_{n,k}^+$, we compute $k^*(\beta_{\min}) \triangleq \min\{k : s_k^{-1}(s_\delta^k) \ge \beta_{\min}\}$.

Letting $s_\delta^{k^*,l}$ be the critical value of $M_{n,k^*,l}^+$, we then calculate $l^* \triangleq \min\{l : s_l^{-1}(s_\delta^{k^*,l}) \geq \beta_{\max}\}$. For the β-CVaR setting, the index k is determined by substituting $\beta_{\min}$ with β in this procedure. This establishes the complete parameter selection mechanism for these methods.

B.2 Conformal Distortion Risk Control via L-statistics

We next introduce conformal distortion risk control via L-statistics (**CDRC-L**) [12]. The distortion risk measure serves as an alternative terminology for the quantile-based risk measure. Given a calibration set $X_{1:n} = \{X_1, \ldots, X_n\}$, we arrange the samples into order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$. Subsequently, we construct the empirical risk $\hat{R}_\psi = \sum_{i=1}^n \big[\psi(\frac{i}{n}) - \psi(\frac{i-1}{n})\big] X_{(i)}$. The principal conclusion by Chen et al. [12] demonstrates that, by characterizing the sampling uncertainty of the empirical risk as $\hat{\sigma}^2 = \frac{1}{n^2}\sum_{i,j=1}^n \psi'\big(\frac{i}{n}\big)\,\psi'\big(\frac{j}{n}\big)\,\big(\frac{i\wedge j}{n} - \frac{ij}{n^2}\big)$, the following holds

$$\frac{\sqrt{n}(\hat{R}_\psi - R_\psi)}{\hat{\sigma}^2} \xrightarrow{d} N(0,1).$$

We then define the upper bound $\hat{R}_\psi^+ = \hat{R}_\psi + z_{1-\delta}\frac{\hat{\sigma}}{\sqrt{n}}$, which guarantees the asymptotic result

$$\liminf_{n\to\infty} P_D\big(R_\psi \leq \hat{R}_\psi^+\big) \geq 1 - \delta.$$

Here, $z_{1-\delta}$ denotes the $(1-\delta)$ quantile of the standard normal distribution $N(0,1)$. Note that this theoretical property fundamentally requires the size of the calibration set to approach infinity. Consequently, the risk control behavior remains indeterminate for finite sample sizes, implying that the guarantee might fail when the number of samples is insufficiently large.

C Algorithms of OQRC

In this section, we present two algorithms: one for deriving the risk bounds of OQRC via Monte-Carlo sampling (Algorithm 1), and another for selecting λ to achieve QRC (Algorithm 2).

Algorithm 1 Monte-Carlo Occupancy-based Quantile Risk Control (OQRC)

Require: Calibration loss samples $X_{1:n}$, test set size m, confidence level δ, bounds U, weighting function ψ, Monte-Carlo replicates B.
Ensure: Upper bound r^* for the test risk measure at confidence level $1-\delta$.

$(X_{(1)}, X_{(2)}, \ldots, X_{(n)}) \leftarrow \mathrm{Sort}(X_{1:n})$
$X_{(n+1)} \leftarrow U$
for $b = 1$ **to** B **do**
  $\mathcal{T} \leftarrow \mathrm{SampleWithoutReplacement}(\{1, 2, \ldots, m+n\}, n)$
  $(T_1, T_2, \ldots, T_n) \leftarrow \mathrm{Sort}(\mathcal{T})$
  $T_0 \leftarrow 0, \quad T_{n+1} \leftarrow m+n+1$
  $S_{-1} \leftarrow 0$
  for $i = 0$ **to** n **do**
    $N_i^{(b)} \leftarrow T_{i+1} - T_i - 1$
    $S_i \leftarrow S_{i-1} + N_i^{(b)}$
  end for
  $\Phi^+(\mathbf{x}^{(b)}; \lambda) \leftarrow \sum_{i=0}^n \int_{\frac{S_{i-1}(\mathbf{x}^{(b)})}{m}}^{\frac{S_i(\mathbf{x}^{(b)})}{m}} \psi(t) X_{(i+1)}\, dt$
end for
$r^* \leftarrow \mathrm{EmpiricalQuantile}\Big(\big\{\Phi^+(\mathbf{x}^{(b)}; \lambda)\big\}_{b=1}^B, 1-\delta\Big)$
return r^*

D Tightness of OQRC

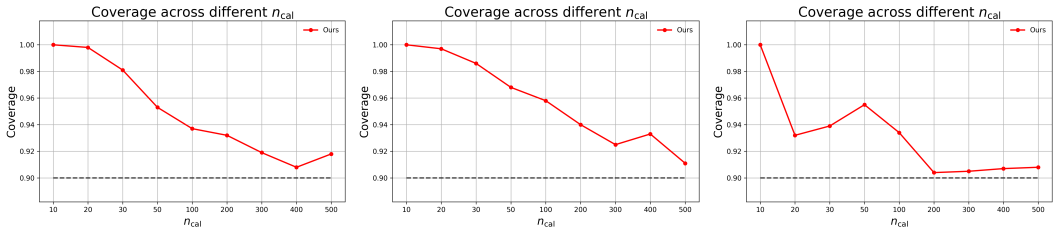
In this appendix, we discuss why the coverage in Table 1 deviates from $1-\delta$. This deviation occurs because FNP is not a continuous loss function and fails to satisfy the condition that f_X is bounded

Algorithm 2 Calibrating λ for Guaranteed Risk Control via OQRC

Require: Calibration samples $X_{1:n}$, test size m, target risk level α, confidence level δ, bounds U, weighting function ψ, candidate set $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$, Monte-Carlo replicates B
Ensure: Calibrated parameter λ^* satisfying the OQRC risk guarantee.

$\Lambda \leftarrow \mathrm{sort}_{\uparrow}(\Lambda)$
for $k = 1$ **to** K **do**
  $X_{1:n}^{\lambda_k} \leftarrow \{X_1^{\lambda_k}, \ldots, X_n^{\lambda_k}\}$
  $r_k \leftarrow$ OQRC-MC$(X_{1:n}^{\lambda_k}, m, \delta, U, \psi, B)$ {using Algorithm 1}
  if $r_k \leq \alpha$ **then**
    $\lambda^* \leftarrow \lambda_k$
    return λ^*
  end if
end for
return FAIL

from below. Our tightness theory relies on the assumption that the loss function possesses such ideal properties. To empirically validate our theory, we construct a synthetic dataset containing 5000 points. The loss value for each point is sampled randomly between 0 and 1 when $\lambda = 0$. It then follows a linear function with respect to λ and decreases to zero at $\lambda = 1$. Figure 5 demonstrates that the coverage converges to $1 - \delta$ under both risk measures as the size of the calibration set increases. We leave the establishment of the OQRC tightness theory for more general loss functions to future work.



(a) $[0.85, 0.95]$-VaR-Interval (b) 0.9-CVaR (c) 0.8-VaR

Figure 5: Coverage in synthetic dataset across different calibration set size.

E Implementation details

E.1 Loss function details

We use the false negative proportion (FNP) as our loss function, and we introduce FNP below. In the context of a multi-label classification problem, consider a model $h : \mathcal{Z} \to [0,1]^k$, where k denotes the number of classes. Let the prediction set generated by the model be defined as $\mathcal{C}_\lambda(Z) = \{h(Z)_y \geq 1 - \lambda\}$, with λ serving as the tuning parameter. Letting Y denote the set of true labels associated with Z, the FNP can be formulated as

$$l_\lambda^{\mathrm{FNP}}(Z) = 1 - \frac{|Y \cap \mathcal{C}_\lambda(Z)|}{|Y|}.$$

Tasks such as tumor segmentation and text emotion recognition can also be formulated in a similar manner. Specifically, the output set for tumor segmentation comprises the pixels within a medical image, whereas the output set for text emotion recognition constitutes a set of emotion labels. It can be shown that the FNP defined in this way is non-increasing with respect to λ.

E.2 Training details

All models are trained on an NVIDIA GeForce RTX 4090 GPU. For the tumor segmentation task, we utilized a training set comprising 1,450 images strictly disjoint from the validation set. We employed a combination of binary cross-entropy and weighted Intersection over Union (IoU) loss as the loss

function. The training process spanned 20 epochs with a batch size of 16, optimized via the Adam optimizer using a learning rate of $1\mathrm{e}-4$.

For the image multi-label classification task, we use the resnet50 model for ResNet, tf_efficientnet_b3 for EfficientNet, tresnet_m for TResNet, and convnext_base for ConvNeXt. We train the models on 118,287 images. The loss function for all these models is uniformly set to BCEWithLogitsLoss, and the optimizer is uniformly AdamW with a weight decay of 1e-4. The learning rate is set to 1e-4 with a CosineAnnealingLR scheduler. All models are trained for 5 epochs with a batch size of 64.

In the text emotion recognition task, we utilized the pre-fine-tuned bert-base-go-emotion model directly from the Hugging Face repository[3].

E.3 Evaluation metrics

We evaluate the performance of risk control framework using three specific metrics. The first metric is the coverage (**Cov**), which represents the empirical probability that the risk on the test set remains under a predefined risk level α. We define this as

$$\text{Cov} = \frac{1}{n_{\text{trials}}} \sum_{i=1}^{n_{\text{trials}}} \mathbf{1}_{R_{\text{test},i} \leq \alpha}.$$

A method is considered valid if the coverage exceeds $1-\delta$, and invalid otherwise. The second metric is the average risk gap (**RiskGap**), which assesses the conservatism of the method and is defined as

$$\text{RiskGap} = \frac{1}{n_{\text{trials}}} \sum_{i=1}^{n_{\text{trials}}} |\alpha - R_{\text{test},i}|.$$

Assuming that the method is valid, a larger risk gap indicates a more conservative approach, which implies that a smaller gap corresponds to tighter control. Finally, we utilize the average prediction set size (**AvgSize**) to measure the magnitude of the generated prediction sets, which is defined as

$$\text{AvgSize} = \frac{1}{n_{\text{trials}}} \sum_{i=1}^{n_{\text{trials}}} \Big(\frac{1}{m} \sum_{Z_{\text{test}} \in \mathcal{D}_{\text{test},i}} |\mathcal{C}_{\lambda_i}(Z_{\text{test}})|\Big).$$

Similar to the risk gap, when the method is valid, a larger average size denotes a higher level of conservatism, whereas a smaller size signifies tight control of the risk.

F Asymptotic Behavior in the Large Test-Set Regime

We now discuss the behavior of OQRC in the large test-set regime, where the number of test losses m tends to infinity while the calibration set size n remains fixed. In this setting, the empirical distribution $\widehat{F}_m$ of the test losses converges to the loss distribution F_X, and the corresponding empirical risk $R_\psi(\widehat{F}_m;\lambda)$ approaches the risk $R_\psi(F_X;\lambda)$. Therefore, the objective of OQRC changes from controlling risk on finite test set to controlling risk over the entire loss distribution.

In this context, we draw n independent uniform random variables on $[0,1]$, and sort them as $0 < U_{(1)} < \cdots < U_{(n)} < 1$, and we define the spacings

$$q_0 = U_{(1)}, \qquad q_i = U_{(i+1)} - U_{(i)}, \quad 1 \leq i \leq n-1, \qquad q_n = 1 - U_{(n)}.$$

The resulting vector $q = (q_0, \ldots, q_n)$ follows the uniform distribution over a simplex, that is

$$q \sim \text{Dirichlet}(1, \ldots, 1).$$

We then define $\Phi^+_\infty(q;\lambda, X_{1:n}) = \sum_{i=0}^{n} \int_{U_{(i)}}^{U_{(i+1)}} \psi(t) X_{(i)} dt$, where we set $U_{(n+1)} = 1$. Consequently, in the large-test-set regime, the distribution F^+ of the upper risk is replaced by

$$F^+_\infty(r;\lambda, X_{1:n}) = \mathbb{P}_{q\sim\text{Dirichlet}(1,\ldots,1)}\left(\Phi^+_\infty(q;\lambda, X_{1:n}) \leq r\right).$$

Subsequently, we define $R^+_\infty = (F^+_\infty)^{-1}(1-\delta)$. As a result, R^+_∞ is a UCB of $R_\psi(\widehat{F}_m;\lambda)$.

[3] https://huggingface.co/bhadresh-savani/bert-base-go-emotion

For unconditional risk control, this limiting representation can be summarized as

$$\mathbb{E}\left[R_\psi(\widehat{F}_m;\lambda)\right] = \mathbb{E}_{q\sim\mathrm{Dirichlet}(1,\ldots,1)}\left[\Phi^+_\infty(q;\lambda,X_{1:n})\right].$$

The Monte-Carlo implementation in this regime is correspondingly simplified. When $m\to\infty$, each Monte-Carlo replicate samples a continuous simplex vector

$$q^{(b)} \sim \mathrm{Dirichlet}(1,\ldots,1), \qquad b=1,\ldots,B,$$

and computes $\Phi^+_\infty(q^{(b)};\lambda,X_{1:n})$. The resulting empirical distribution

$$\widehat{F}^+_{\infty,B}(r;\lambda,X_{1:n}) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\left\{\Phi^+_\infty(q^{(b)};\lambda,X_{1:n})\le r\right\}$$

approximates $F^+_\infty(r;\lambda,X_{1:n})$. The empirical $(1-\delta)$-quantile of these Monte-Carlo values gives the UCB of the quantile risk, while their empirical average,

$$\widehat{\overline{R}}^+_{\infty,B}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\Phi^+_\infty(q^{(b)};\lambda,X_{1:n}),$$

estimates the $\mathbb{E}\left[R_\psi(\widehat{F}_m;\lambda)\right]$.

G How OQRC behave under distribution shift

To examine the behavior of OQRC when Assumption 2.1 is violated, we construct two controlled distribution shifts on MS-COCO: a covariate shift induced by image brightness and a class-frequency shift induced by category rarity. In both settings, OQRC is applied exactly as in the i.i.d. case, i.e., without importance weighting and without any knowledge of the shift.

We use $N = 2000$ images of the MS-COCO 2014 validation split, each annotated with a subset $y \subseteq \{1,\ldots,K\}$ of $K = 80$ object categories. Image-level class scores are obtained from the COCO-pretrained Faster R-CNN with a ResNet-50-FPN backbone released in `torchvision`. For each category, we take the maximum confidence over all detections of that category. The resulting score vector of each image is then min–max normalized to $[0,1]$. For a threshold parameter λ, the prediction set is $\mathcal{C}_\lambda(x) = \{k : s_k(x) > 1-\lambda\}$, where λ ranges over a grid Λ of 1000 equally spaced values in $[0,1]$, so that larger values of λ yield larger prediction sets. We adopt the false negative proportion (FNP) loss. The loss is bounded in $[0,1]$ and non-increasing in λ.

Both shifts are generated by exponentially tilting the sampling distribution over the pool $\mathcal{D} = \{(x_i,y_i)\}_{i=1}^N$ according to a scalar shift score c_i. To make the shift strength comparable across scores of different scales, we first map the scores to normalized ranks $u_i = -1+2(\rho_i-1)/(N-1) \in [-1,1]$, where ρ_i is the rank of c_i among $c_1,\ldots,c_N$, and ties are broken by index. The calibration and test sampling distributions are tilted in opposite directions,

$$p_i^{\mathrm{cal}} \propto \exp(-\gamma u_i), \qquad p_i^{\mathrm{test}} \propto \exp(\gamma u_i),$$

with shift strength $\gamma = 2$. The likelihood ratio $p_i^{\mathrm{test}}/p_i^{\mathrm{cal}} = \exp(2\gamma u_i)$ lies in $[e^{-4},e^4]$. Hence, the two distributions share the same support but concentrate their mass at opposite ends of the shift score. In each trial, we draw $n = 200$ calibration points from $\mathcal{D}$ without replacement with probabilities proportional to p^{cal}, and then draw $m = 500$ test points without replacement from the remaining $N-n$ points with probabilities proportional to p^{test}, so that the calibration and test sets are disjoint. This procedure is repeated for $T = 1000$ independent trials.

In the covariate shift setting, the shift score is the mean brightness of the image, computed as the average pixel intensity over the three RGB channels of a downsampled copy of the image and rescaled to $[0,1]$. Because the sampling weights depend on an example only through its image, the selection is independent of the labels given the image. Consequently, $P_{\mathrm{cal}}(Y \mid X) = P_{\mathrm{test}}(Y \mid X)$, and only the marginal distribution of X changes.

In the class-frequency shift setting, let $\pi_k = N^{-1}\sum_{i=1}^N \mathbf{1}\{k\in y_i\}$ denote the prevalence of category k in the pool, and define the rarity of category k as $r_k = -\log\max(\pi_k, 1/N)$. The shift score of an

image is the average rarity of its annotated categories,

$$c_i = \frac{1}{|y_i|} \sum_{k \in y_i} r_k,$$

with $c_i = 0$ for the 14 images without annotations. Under this construction, the calibration data over-represent images containing frequent categories, whereas the test data over-represent images containing rare categories. Since the sampling weights depend on an example only through its labels, $P_{\text{cal}}(X \mid Y) = P_{\text{test}}(X \mid Y)$. That is, the induced shift is a label shift rather than a covariate shift.

Table 3 indicates that both covariate shift and class-frequency shift degrade OQRC. Under covariate shift, OQRC remains valid but becomes overly conservative. Under Class-frequency shift, OQRC underestimates the test risk and selects an overly small $\hat{\lambda}$.

Table 3: Performance under Different Distribution Shifts.

Metrics	Risk Measure	No Shift	Covariate Shift	Class-frequency Shift
Coverage	VaR-Interval	0.904	0.978	0.491
	CVaR	0.959	0.997	0.689
	VaR	0.922	0.999	0.670
RiskGap	VaR-Interval	0.066 ± 0.041	0.087 ± 0.045	0.042 ± 0.033
	CVaR	0.082 ± 0.050	0.143 ± 0.058	0.051 ± 0.038
	VaR	0.058 ± 0.028	0.068 ± 0.017	0.052 ± 0.038
AvgSize	VaR-Interval	3.115 ± 0.237	3.945 ± 0.194	2.514 ± 0.196
	CVaR	3.345 ± 0.255	4.231 ± 0.256	2.663 ± 0.203
	VaR	2.417 ± 0.109	3.316 ± 0.082	1.959 ± 0.088

H Additional experimental results

H.1 Extended results across different test size m

Table 4 presents the variation of coverage, RiskGap, AvgSize, and runtime across different test sizes m. We conduct experiments using various risk measures on MS COCO with ResNet50 (other experimental parameters remain consistent with the default settings in the original manuscript). The results indicate that all evaluated metrics exhibit substantial variation across different test size. Therefore, we recommend that practitioners set m to the actual test size whenever it is known in advance. Otherwise, practitioners should either estimate the test size or adopt the formulation in Appendix F, which addresses the case in which the test size tends to infinity.

H.2 Extended results across different Monte Carlo occupancy samples

Our method is insensitive to the number of Monte Carlo samples, especially when the number is sufficiently large (e.g., 1000). To validate this, we conduct experiments using various risk measures on MS COCO with ResNet50 (other experimental parameters remain consistent with the default settings in the original manuscript) and the results are presented in Table 5. The results indicate that all metrics stabilize once the sample count is sufficiently large, whereas larger sample counts lead to more stable performance before then. In addition, the computational burden is acceptable even with a large number of samples. From a practical perspective, we recommend setting the number of Monte Carlo samples to at least 10,000.

H.3 Extended results across diverse hyperparameter configurations

Figures 6 to 14 compare our method with the baseline approaches in terms of Cov, RiskGap, and AvgSize across three tasks under three risk measures. Within each figure, the top three subplots

Table 4: Performance comparison under different sample sizes.

Metrics	Risk Measure	$m = 10$	$m = 50$	$m = 100$	$m = 500$	$m = 1000$
Coverage	VaR-Interval	0.910	0.899	0.903	0.904	0.901
	CVaR	0.946	0.942	0.954	0.959	0.955
	VaR	0.910	0.918	0.928	0.922	0.931
RiskGap	VaR-Interval	0.263 ± 0.126	0.126 ± 0.087	0.097 ± 0.063	0.066 ± 0.041	0.063 ± 0.040
	CVaR	0.261 ± 0.128	0.146 ± 0.085	0.117 ± 0.070	0.082 ± 0.050	0.073 ± 0.044
	VaR	0.264 ± 0.126	0.127 ± 0.084	0.101 ± 0.071	0.058 ± 0.028	0.055 ± 0.045
AvgSize	VaR-Interval	4.804 ± 1.345	3.392 ± 0.406	3.261 ± 0.299	3.115 ± 0.237	3.103 ± 0.226
	CVaR	4.821 ± 1.336	3.659 ± 0.484	3.500 ± 0.353	3.345 ± 0.255	3.303 ± 0.230
	VaR	3.786 ± 1.372	2.684 ± 0.473	2.399 ± 0.291	2.417 ± 0.109	2.324 ± 0.230
RunTime (s)	VaR-Interval	0.107	0.126	0.138	0.156	0.187
	CVaR	0.110	0.127	0.139	0.158	0.187
	VaR	0.096	0.114	0.126	0.145	0.175

Table 5: Performance comparison under different values of B.

Metrics	Risk Measure	$B = 100$	$B = 200$	$B = 500$	$B = 1000$	$B = 10000$	$B = 100000$
Coverage	VaR-Interval	0.895	0.896	0.902	0.903	0.904	0.903
	CVaR	0.950	0.956	0.953	0.956	0.959	0.960
	VaR	0.900	0.916	0.919	0.916	0.922	0.918
RiskGap	VaR-Interval	0.070 ± 0.043	0.068 ± 0.041	0.071 ± 0.042	0.067 ± 0.041	0.066 ± 0.041	0.066 ± 0.040
	CVaR	0.086 ± 0.052	0.084 ± 0.051	0.083 ± 0.051	0.081 ± 0.053	0.082 ± 0.050	0.080 ± 0.049
	VaR	0.055 ± 0.028	0.058 ± 0.028	0.057 ± 0.029	0.057 ± 0.028	0.058 ± 0.028	0.057 ± 0.029
AvgSize	VaR-Interval	3.134 ± 0.243	3.128 ± 0.231	3.146 ± 0.239	3.117 ± 0.241	3.115 ± 0.237	3.115 ± 0.242
	CVaR	3.362 ± 0.267	3.349 ± 0.258	3.347 ± 0.248	3.346 ± 0.267	3.345 ± 0.255	3.346 ± 0.263
	VaR	2.420 ± 0.115	2.417 ± 0.113	2.414 ± 0.107	2.414 ± 0.114	2.417 ± 0.109	2.418 ± 0.109
RunTime (s)	VaR-Interval	0.004	0.005	0.009	0.017	0.156	1.818
	CVaR	0.004	0.005	0.010	0.017	0.158	1.907
	VaR	0.003	0.005	0.009	0.015	0.145	1.590

display the coverage, the middle three present the RiskGap, and the bottom three illustrate the AvgSize. The experimental results demonstrate that our method consistently maintains the required coverage while significantly reducing both the RiskGap and the AvgSize compared to the baselines across diverse datasets, risk measures, and hyperparameter configurations.

H.4 Extended results across diverse model architectures

Figure 15 compares the performance of our method with the baselines across different model architectures under $[0.85, 0.95]$-VaR-Interval. Figure 16 presents the comparison under 0.9-CVaR. Figure 17 presents the comparison under 0.8-VaR. The hyperparameters match the configurations used for the performance evaluation experiment across model architectures in Section 4. The experimental results demonstrate that our approach significantly narrows both the RiskGap and the AvgSize compared to the baselines while achieving valid coverage.

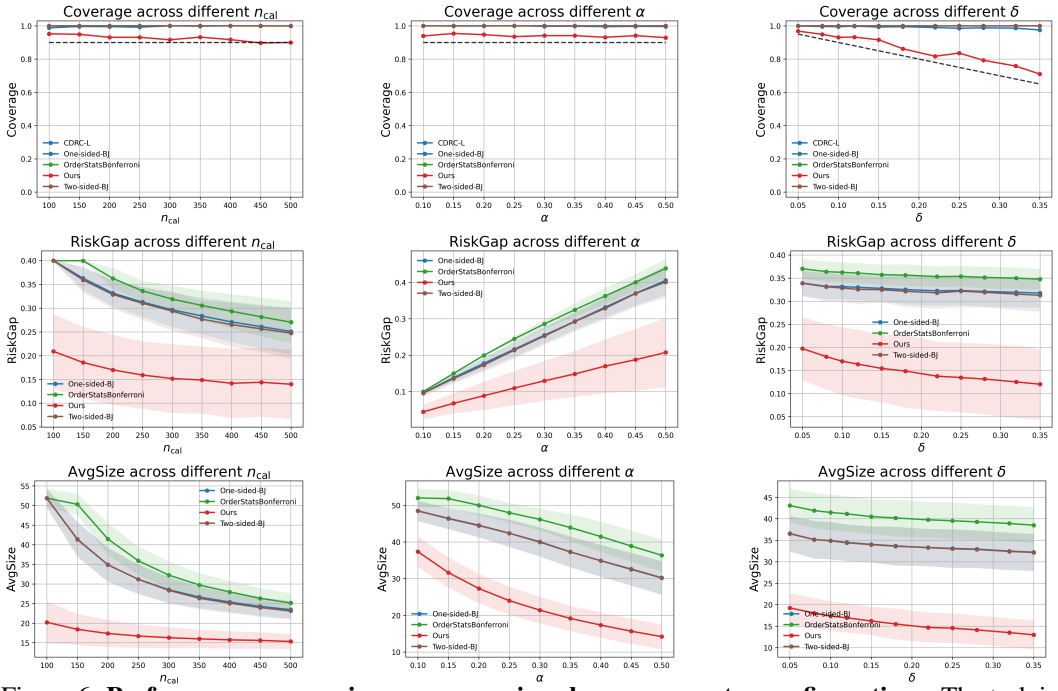


Figure 6: **Performance comparison across various hyperparameter configurations.** The task is tumor segmentation, and the risk measure is $[0.85, 0.95]$-VaR-Interval.

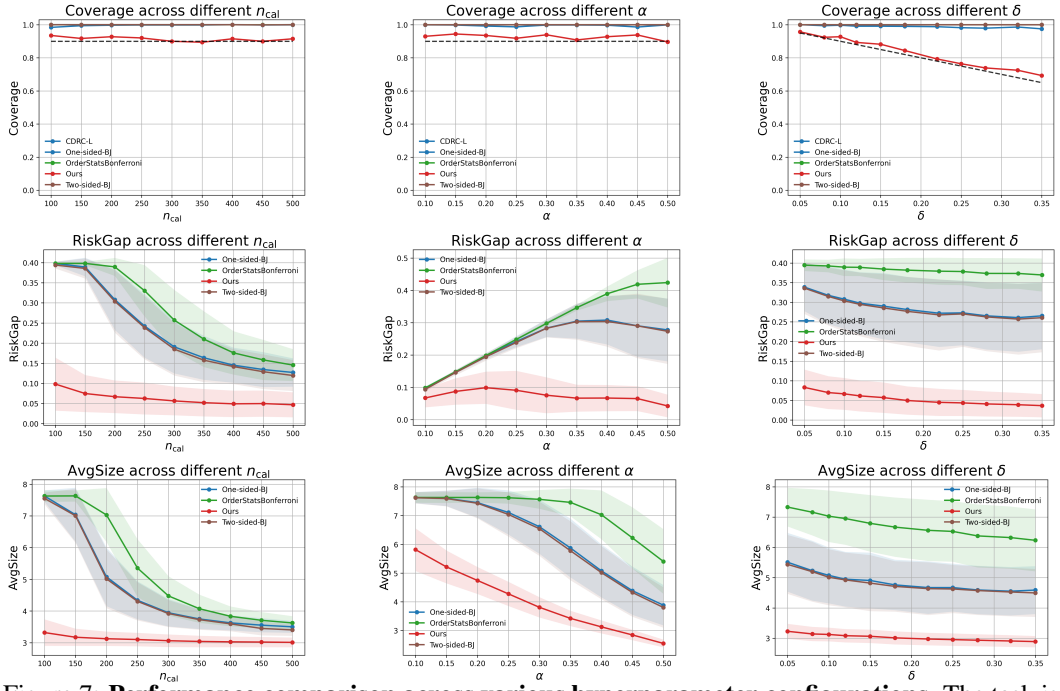


Figure 7: **Performance comparison across various hyperparameter configurations.** The task is image multi-label classification, and the risk measure is $[0.85, 0.95]$-VaR-Interval.

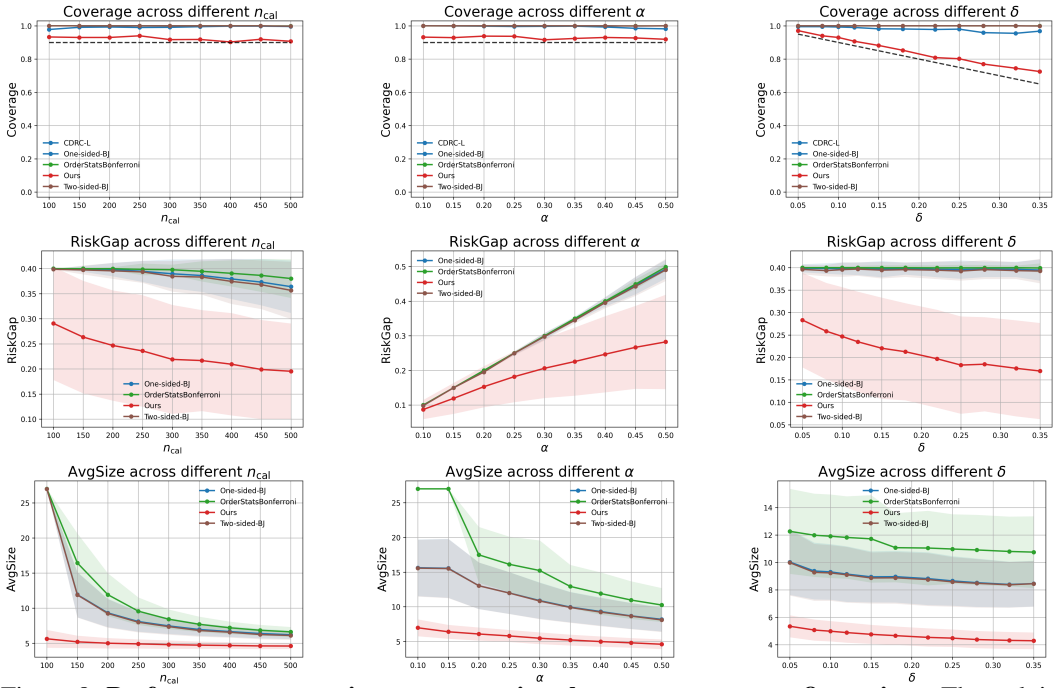


Figure 8: **Performance comparison across various hyperparameter configurations.** The task is text emotion recognition, and the risk measure is $[0.85, 0.95]$-VaR-Interval.

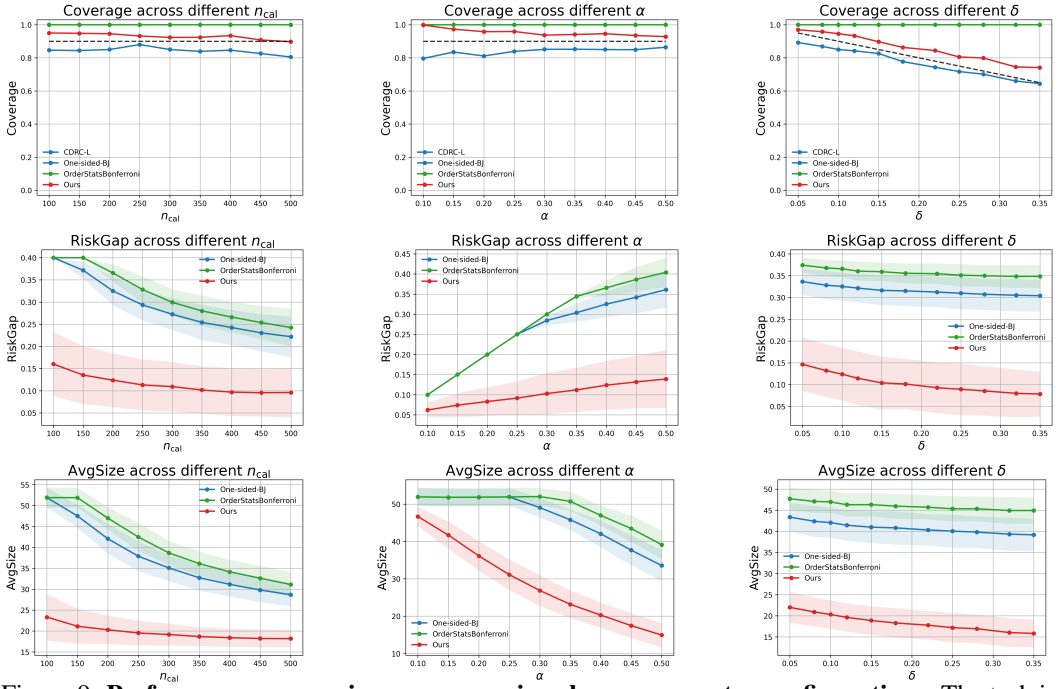


Figure 9: **Performance comparison across various hyperparameter configurations.** The task is tumor segmentation, and the risk measure is 0.9-CVaR.

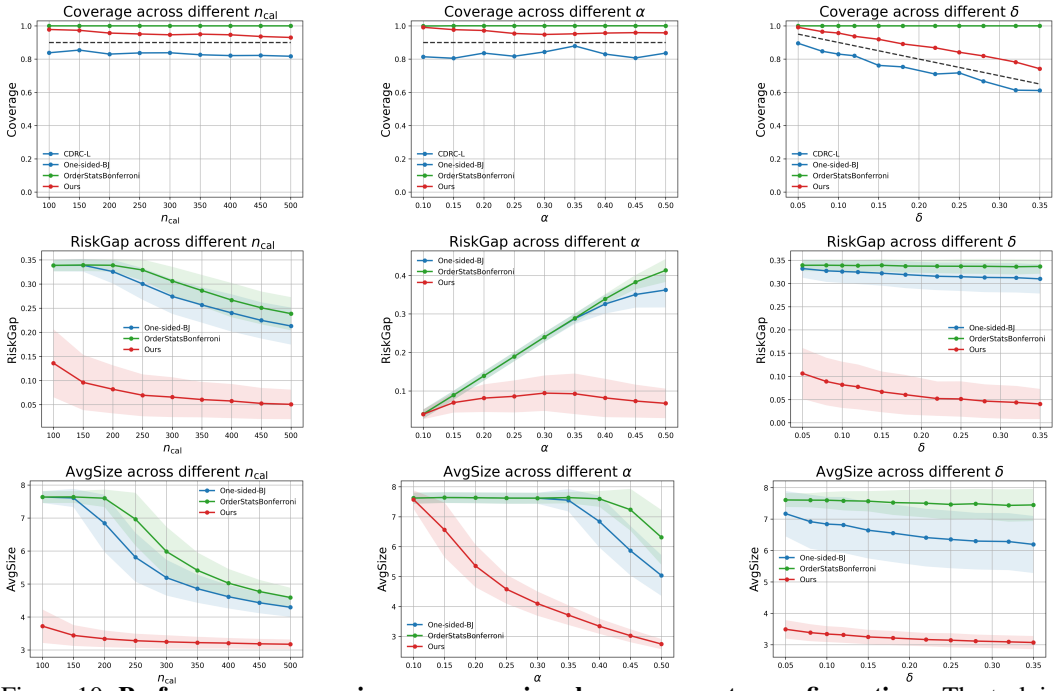


Figure 10: **Performance comparison across various hyperparameter configurations.** The task is image multi-label classification, and the risk measure is 0.9-CVaR.

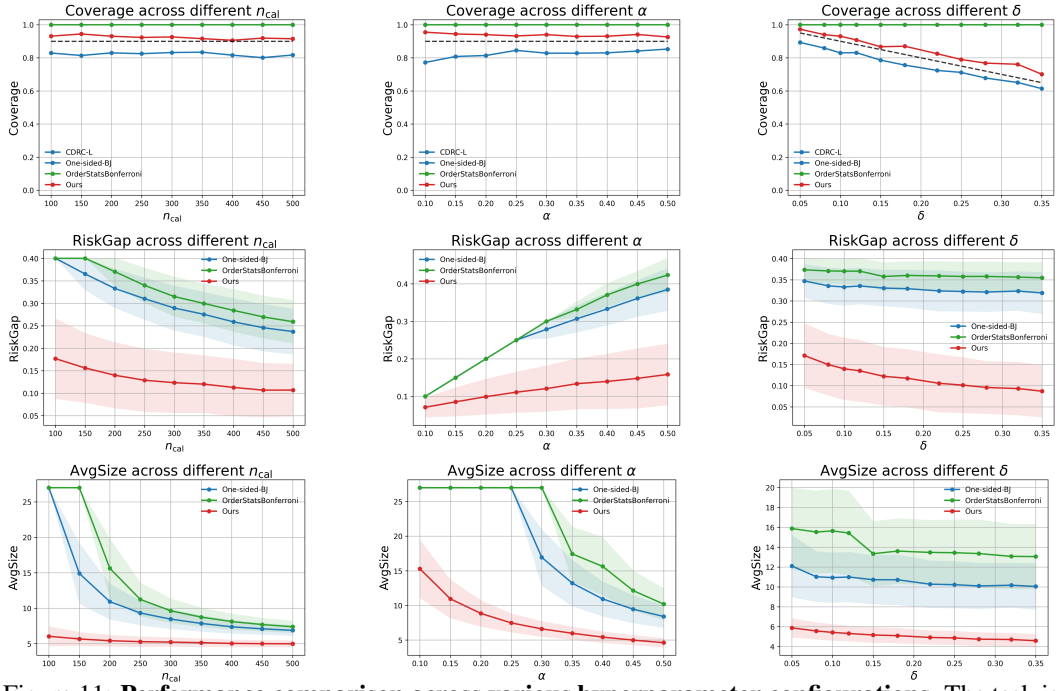


Figure 11: **Performance comparison across various hyperparameter configurations.** The task is text emotion recognition, and the risk measure is 0.9-CVaR.

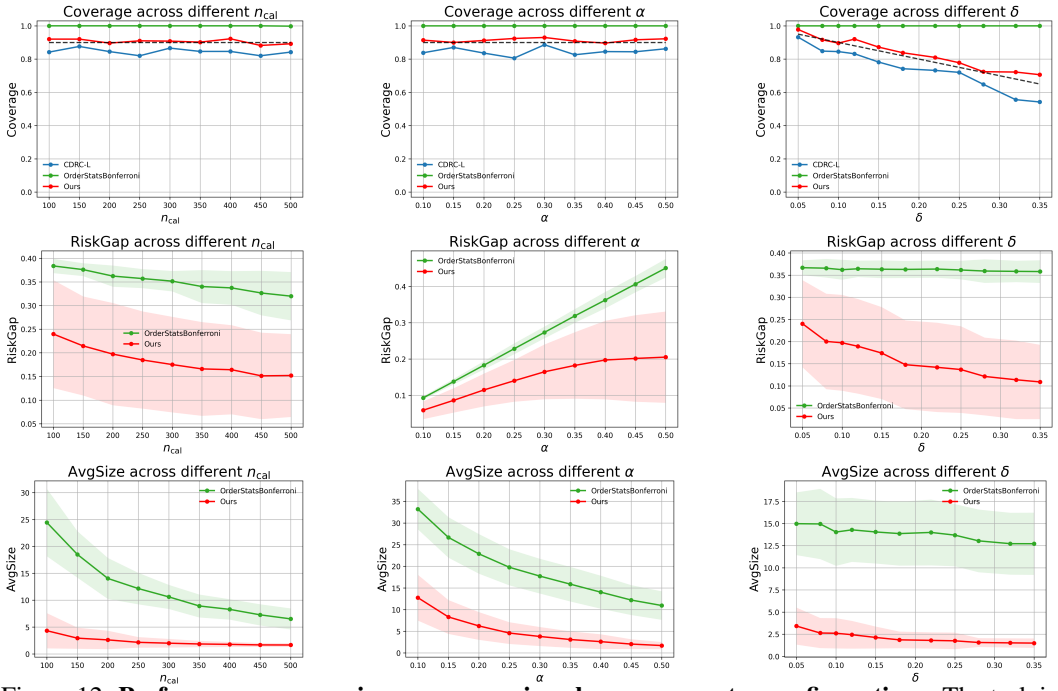


Figure 12: **Performance comparison across various hyperparameter configurations.** The task is tumor segmentation, and the risk measure is 0.8-VaR.

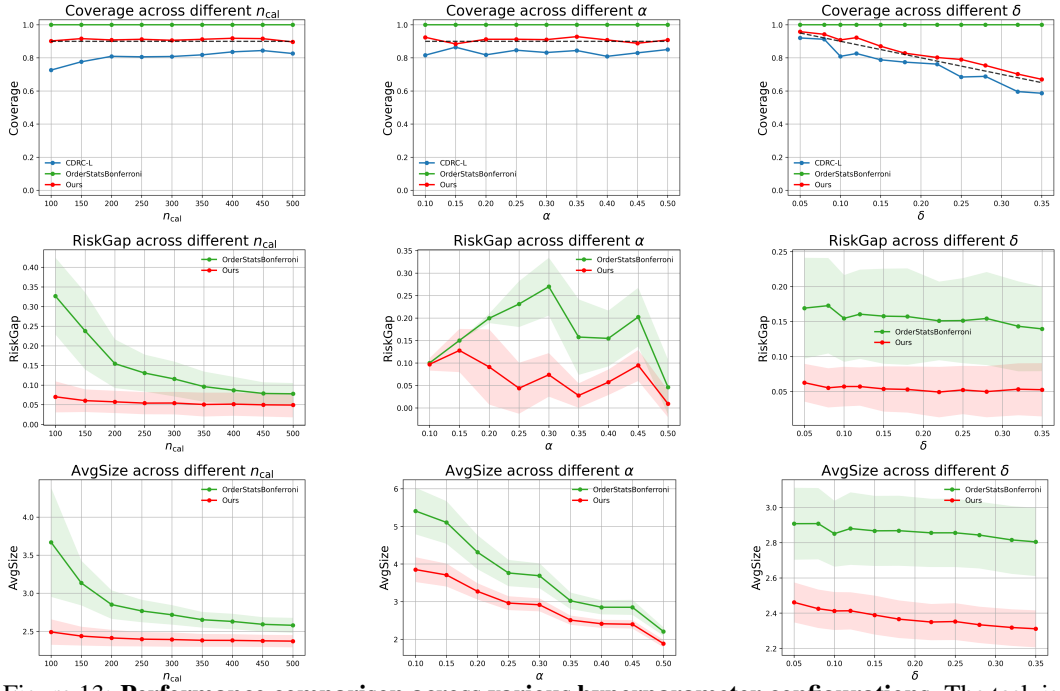


Figure 13: **Performance comparison across various hyperparameter configurations.** The task is image multi-label classification, and the risk measure is 0.8-VaR.

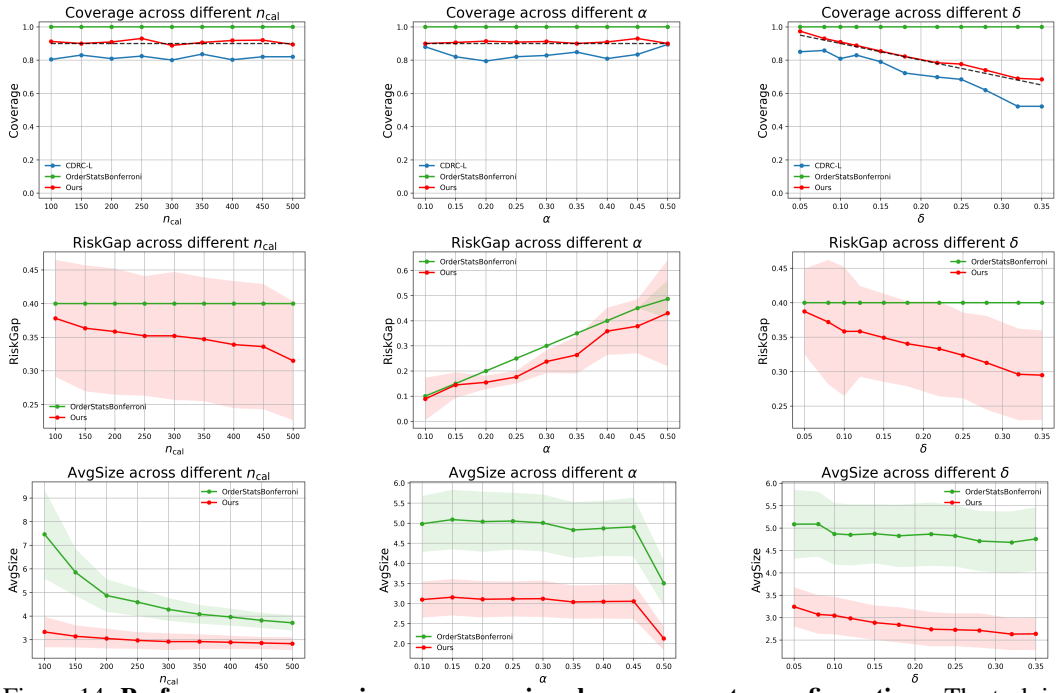


Figure 14: **Performance comparison across various hyperparameter configurations.** The task is text emotion recognition, and the risk measure is 0.8-VaR.

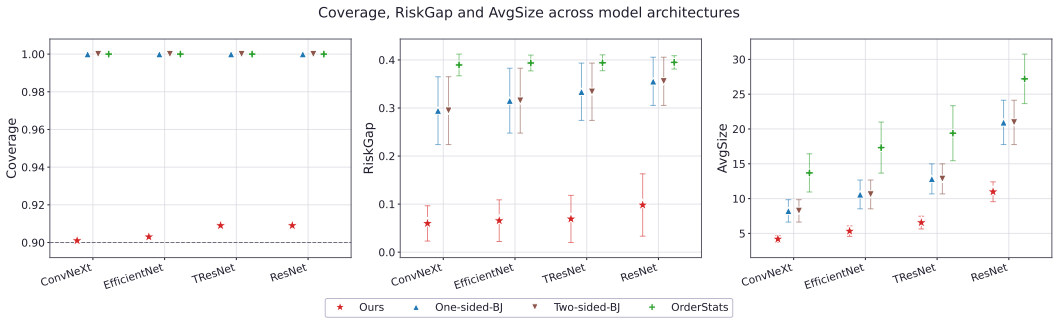


Figure 15: Performance comparison across different model architectures under VaR-Interval.

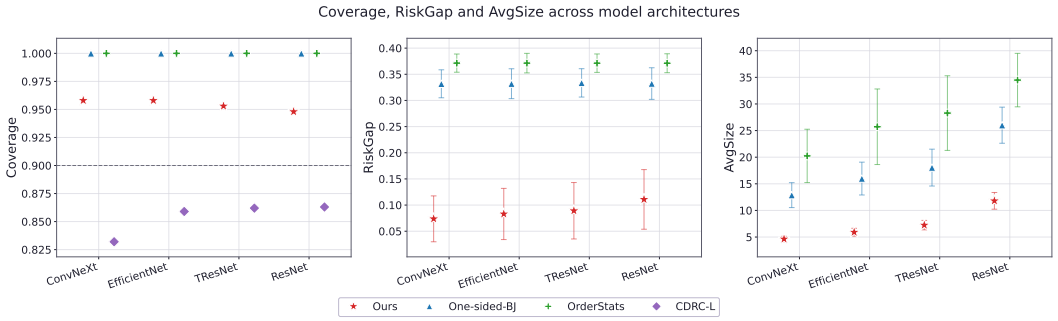


Figure 16: Performance comparison across different model architectures under 0.8-CVaR.

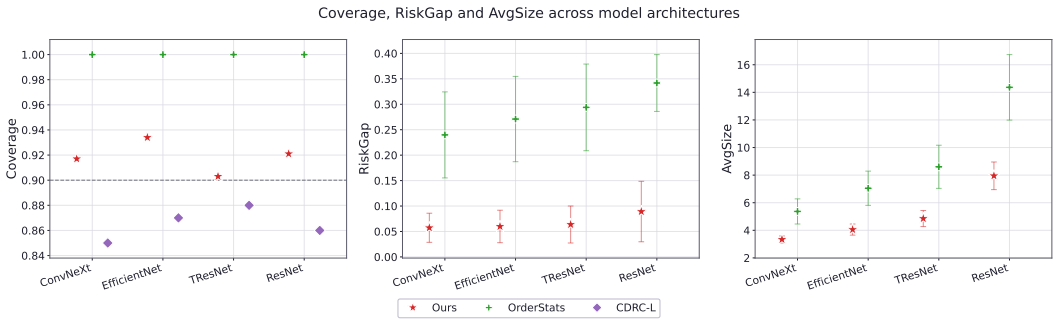


Figure 17: Performance comparison across different model architectures under 0.8-VaR.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: Our theoretical and experimental results consistently support the claims presented in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitation of this work in the conclusion section 7.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: We provide the proof of our theoretical result in the Appendix A.

   Guidelines:

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

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide the detailed implementation details to reproduce our results.

   Guidelines:

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

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

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

   Answer: [Yes]

   Justification: We will release the code of our work.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide the information in Appendix E.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Our results are provided in mean and standard error format.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: The compute resources are provided in Appendix E.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: Yes, we affirm that our research adheres to the NeurIPS Code of Ethics in all aspects, including considerations related to data usage, transparency, and potential societal impact.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: We discuss about the societal impacts of risk control in the introduction.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our work does not involve the release of model and dataset.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The dataset and model we use are both open-sourced. Our code is based on PyTorch.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We will upload our code in the supplement material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.