Spectral Conformal Risk Control: Distribution-Free Tail Guarantees via Bayesian Quadrature

Supplementary Material

Table 6. Notation summary

Symbol	Description
$Z, Z_{1:n}$	Data point(s); calibration sample of size n
$\lambda \in \Lambda$	Control parameter (e.g., threshold)
$\ell(Z;\lambda)$	Deployment loss at Z with parameter λ
$K_\lambda(t)$	Quantile function of loss distribution
ϕ	Spectral density (nonnegative, nondecreasing, $\int_0^1 \phi = 1$)
$\rho_\phi(\lambda)$	Spectral risk $\int_0^1 K_\lambda(t)\phi(t)\,\mathrm{d}t$
$T_{(i)}$	Order statistics of the PIT levels $U_i := F_\lambda(\ell_i(\lambda))$ (distributed as order statistics of n i.i.d. $\mathrm{Unif}(0,1)$)
W_i	ϕ-mass spacing $\int_{T_{(i-1)}}^{T_{(i)}} \phi(t)\,\mathrm{d}t$
$L_\phi^+(\lambda)$	Upper envelope $\sum_{i=1}^{n+1} W_i \ell_{(i)}(\lambda)$

A. Mathematical Background

A.1. Probability, Quantiles, and Law-Invariant Coherent Risks

Probability spaces and L^p. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $\mathcal{F}$ a σ-algebra and $\mathbb{P}$ a probability measure. We say it is *atomless* if every $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ contains a subset of strictly smaller positive probability. For $1 \le p \le \infty$, write $L^p = L^p(\Omega, \mathcal{F}, \mathbb{P})$; equalities/inequalities are interpreted $\mathbb{P}$-a.s.

Distribution and quantiles (loss convention). For a real r.v. X, write $F_X(x) = \mathbb{P}(X \le x)$ (right-continuous). We use the left-continuous generalized inverse

$$F_X^{-1}(u) := \inf\{x \in \mathbb{R} : F_X(x) \ge u\}, \qquad u \in (0,1),$$

and adopt the *loss convention* (larger X means worse). This pins down $\mathrm{VaR}_u(X) := F_X^{-1}(u)$ at atoms.

VaR and CVaR. For $X \in L^1$ and $\alpha \in [0,1)$,

$$\begin{aligned}\mathrm{CVaR}_\alpha(X) &= \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(X)\,\mathrm{d}u \\ &= \min_{t\in\mathbb{R}} \Big\{ t + \tfrac{1}{1-\alpha}\,\mathbb{E}[(X-t)_+] \Big\}.\end{aligned}$$

(The variational identity is due to [33].)

Coherent, law-invariant, comonotone-additive. A monetary risk measure $\rho : \mathcal{X} \to \mathbb{R}$ (with $\mathcal{X} \subseteq L^p$) is *coherent* if it satisfies monotonicity, subadditivity, positive homogeneity, and cash/translation invariance $\rho(X + m) = \rho(X) + m$ (loss convention). It is *law-invariant* if $X \overset{d}{=} Y \Rightarrow \rho(X) = \rho(Y)$. Comonotone-additivity means $\rho(X+Y) = \rho(X) + \rho(Y)$ whenever X, Y are comonotone.

Comonotonicity. Random variables X, Y are *comonotone* iff there exist a r.v. Z and nondecreasing f, g with $X = f(Z)$, $Y = g(Z)$ a.s.; equivalently $(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \ge 0$ for $\mathbb{P}\otimes\mathbb{P}$-a.s. (ω, ω').

Distortion (Choquet) risks. Let $g : [0,1] \to [0,1]$ be nondecreasing with $g(0) = 0, g(1) = 1$. For integrable X, define

$$\rho_g(X) = \int_{-\infty}^0 \big(g(\mathbb{P}(X \ge x)) - 1\big)\,\mathrm{d}x + \int_0^\infty g(\mathbb{P}(X \ge x))\,\mathrm{d}x.$$

(Some authors use $>$ rather than $\ge$ in $\bar{F}_X(x) = \mathbb{P}(X \ge x)$; the difference is immaterial unless at atoms.) If g is concave, then ρ_g is coherent and comonotone-additive (Choquet integral of a concave distortion); see, e.g., [42] for a modern treatment.

Spectral risks and mixtures of CVaR. A *spectral* risk measure is indexed by a nonnegative, nondecreasing, normalized $\phi \in L^1([0,1])$:

$$\rho_\phi(X) = \int_0^1 \mathrm{VaR}_u(X)\,\phi(u)\,\mathrm{d}u.$$

PIT levels and order statistics. For each λ, define the probability integral transform (PIT) levels $U_i := F_\lambda(\ell_i(\lambda))$ from the calibration losses; under (A1)–(A2) these satisfy $U_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$. Let $T_{(1)} \le \cdots \le T_{(n)}$ denote the order statistics of $(U_1, \dots, U_n)$; their joint law is that of order statistics of n i.i.d. uniform draws, but they remain *dependent on the calibration sample*. We write $T_{(0)} := 0$ and $T_{(n+1)} := 1$, and define $W_i := \int_{T_{(i-1)}}^{T_{(i)}} \phi(t)\,\mathrm{d}t$ so that $L_\phi^+(\lambda) = \sum_{i=1}^{n+1} W_i\, \ell_{(i)}(\lambda)$ uses the same PIT-derived spacings. Every spectral risk is a *mixture of CVaRs*: there exists a probability measure μ on $[0,1)$ such that

$$\begin{aligned}\rho_\phi(X) &= \int_0^1 \mathrm{CVaR}_\alpha(X)\,\mu(\mathrm{d}\alpha), \\ \phi(u) &= \int_{[0,u]} \frac{1}{1-\alpha}\,\mu(\mathrm{d}\alpha) \quad \text{for } u \in [0,1).\end{aligned}$$

This representation is due to [1] and [2].

Kusuoka representation. Assume $(\Omega, \mathcal{F}, \mathbb{P})$ is atomless and $\rho : L^p \to \mathbb{R}$ is law-invariant and coherent with the Fatou (Lebesgue) property. Then there exists a nonempty family $\mathcal{M}$ of probability measures on $[0, 1]$ such that

$$\rho(X) = \sup_{\mu \in \mathcal{M}} \int_0^1 \mathrm{CVaR}_\alpha(X)\, \mu(\mathrm{d}\alpha), \qquad X \in L^p\,.$$

If in addition ρ is comonotone-additive, the supremum reduces to a single μ; hence ρ is spectral. For $p < \infty$, one can (and typically does) take $\mu(\{1\}) = 0$ so finiteness holds for all $X \in L^p$. [21]; see [37] for an expository treatment.

B. Proofs

In this section, we provide the proofs for all results in the main paper.

B.1. Auxillary Lemmas

Lemma B.1 (Measurability)**.** *Fix $\lambda \in \Lambda$ and assume $\phi \in L^1([0, 1])$. Then $L^+_\phi(\lambda)$ is a real-valued random variable, and its $(1 - \delta)$-quantile $Q_{1-\delta}(\lambda)$ is well-defined for every $\delta \in (0, 1)$.*

Proof. The order statistics $(T_{(1)}, \ldots, T_{(n)})$ lie in the simplex $\{0 \le t_{(1)} \le \cdots \le t_{(n)} \le 1\}$ and admit a joint density. Define $F_\phi(t) = \int_0^t \phi(s)\, \mathrm{d}s$. Then F_ϕ has bounded variation and is continuous (being the integral of an L^1 function), and each spacing $W_i = F_\phi(T_{(i)}) - F_\phi(T_{(i-1)})$ is a continuous function of $(T_{(1)}, \ldots, T_{(n)})$. For fixed λ, with a tie-breaking rule (randomized or lexicographic) for equal calibration losses, the ordered losses $\ell_{(i)}(\lambda)$ are measurable functions of the data. Therefore $L^+_\phi(\lambda) = \sum_i W_i \ell_{(i)}(\lambda)$ is Borel measurable. Quantile maps of real-valued random variables are measurable, so $Q_{1-\delta}(\lambda)$ is well-defined for every $\delta \in (0, 1)$. □

Lemma B.2 (Monotonicity in λ)**.** *Assume (A6). If $\lambda \mapsto \ell(z; \lambda)$ is non-increasing for every z, then $\lambda \mapsto K_\lambda(t)$, $\lambda \mapsto \rho_\phi(\lambda)$, and $\lambda \mapsto Q_{1-\delta}(\lambda)$ are non-increasing. If $\lambda \mapsto \ell(z; \lambda)$ is non-decreasing, the same maps are non-decreasing. Consequently, when $\lambda \mapsto \ell(z; \lambda)$ is non-increasing, the feasible set $\{\lambda \in \Lambda : Q_{1-\delta}(\lambda) \le \alpha\}$ is an up-set in the induced order on Λ_{grid}; when $\lambda \mapsto \ell(z; \lambda)$ is non-decreasing, the feasible set is instead a down-set in that order.*

Proof. Assume first that $\lambda \mapsto \ell(z; \lambda)$ is non-increasing. Fix t and $\lambda_1 \le \lambda_2$. Then $\ell(z; \lambda_1) \ge \ell(z; \lambda_2)$ for every z, so $F_{\lambda_1} \le F_{\lambda_2}$ pointwise and therefore $K_{\lambda_1}(t) \ge K_{\lambda_2}(t)$. Integrating against ϕ yields $\rho_\phi(\lambda_1) \ge \rho_\phi(\lambda_2)$, and applying the same argument to the envelope shows $L^+_\phi(\lambda_1) \ge L^+_\phi(\lambda_2)$ and thus $Q_{1-\delta}(\lambda_1) \ge Q_{1-\delta}(\lambda_2)$. If $\lambda \mapsto \ell(z; \lambda)$ is non-decreasing, the inequalities reverse. Therefore, in the non-increasing case the feasible set is upward closed under the induced order on Λ_{grid}, while in the non-decreasing case it is downward closed. □

Lemma B.3 (Antitonicity of Two-Parameter Control)**.** *Under assumption (A6), $\ell(\cdot; \tau, k)$ is antitone in (τ, k) w.r.t. the partial order $\preceq$. Consequently, for any nonnegative spectral density ϕ, $\rho_\phi(\tau, k)$ and $Q_{1-\delta}(\tau, k)$ are antitone in (τ, k). Hence the feasible set $\{(\tau, k) : Q_{1-\delta}(\tau, k) \le \alpha\}$ is an **up-set** under $\preceq$ (closed upward: if (τ_1, k_1) is feasible and $(\tau_1, k_1) \preceq (\tau_2, k_2)$, then (τ_2, k_2) is also feasible). Consequently the feasible frontier contains both the pure-τ and pure-k optima, and the grid minimizer $(\widehat{\tau}, \widehat{k})$ weakly dominates either single-parameter control: if $\widehat{\tau}_*$ and $\widehat{k}_*$ are the optimal single-parameter choices, then*

$$\mathbb{E}[|S(X; \widehat{\tau}, \widehat{k})|] \le \min\left\{\mathbb{E}[|S(X; \widehat{\tau}_*, 1)|],\, \mathbb{E}[|S(X; 0, \widehat{k}_*)|]\right\}.$$

Proof. Since $\ell(z; \tau, k)$ is non-decreasing in τ and non-increasing in k, increasing τ or decreasing k increases the loss pointwise. Under our partial order $(\tau_1, k_1) \preceq (\tau_2, k_2)$ iff $\tau_1 \ge \tau_2$ and $k_1 \le k_2$, so $(\tau_1, k_1) \preceq (\tau_2, k_2)$ means that (τ_2, k_2) has a smaller threshold and/or a larger k, hence smaller losses pointwise. This implies $K_{(\tau_1, k_1)}(t) \ge K_{(\tau_2, k_2)}(t)$ for all t, and therefore $\rho_\phi(\tau_1, k_1) \ge \rho_\phi(\tau_2, k_2)$ and $Q_{1-\delta}(\tau_1, k_1) \ge Q_{1-\delta}(\tau_2, k_2)$.

If (τ_1, k_1) is feasible (i.e., $Q_{1-\delta}(\tau_1, k_1) \le \alpha$) and $(\tau_1, k_1) \preceq (\tau_2, k_2)$, then $Q_{1-\delta}(\tau_2, k_2) \le Q_{1-\delta}(\tau_1, k_1) \le \alpha$, so (τ_2, k_2) is also feasible. Thus the feasible set is an up-set. The pure-τ and pure-k optima are feasible points in the 2D grid, so the optimal feasible pair cannot have larger set size than either. □

Lemma B.4 (Deterministic bound)**.** *Let X be a real-valued random variable and let $c \in \mathbb{R}$ be constant. If $X \ge c$ almost surely and $\mathbb{P}(X \le a) > 0$ for some $a \in \mathbb{R}$, then $c \le a$. In particular, if c is deterministic and $\mathbb{P}(c \le a) > 0$, then $c \le a$.*

Proof. If $c > a$, then on the event $\{X \le a\}$ we would have $X < c$, contradicting $X \ge c$ almost surely. Hence $\mathbb{P}(X \le a) > 0$ is impossible when $c > a$, so necessarily $c \le a$. □

B.2. Envelope

Theorem 3.1 (Envelope inequality)**.** *Under assumptions (A1)–(A6), for any $\lambda \in \Lambda$, the spectral risk of the induced distribution is bounded above by the upper envelope:*

$$\rho_\phi(\lambda) \;\le\; L^+_\phi(\lambda). \tag{10}$$

Moreover, among all nondecreasing quantile functions K that are compatible with the observed calibration order

constraints (i.e., satisfy $K(T_{(i)}) \le \ell_{(i)}(\lambda)$ for all i), the supremum of $\int_0^1 K(t)\phi(t)\,\mathrm{d}t$ is equal to $L_\phi^+(\lambda)$ and is attained by the piecewise-constant choice

$$K_\lambda^*(t) = \sum_{i=1}^{n+1} \ell_{(i)}(\lambda)\, \mathbb{1}_{(T_{(i-1)},T_{(i)}]}(t). \tag{11}$$

Proof. Let K_λ be any nondecreasing quantile function compatible with the calibration order constraints, i.e., such that $K_\lambda(T_{(i)}) \le \ell_{(i)}(\lambda)$ for all i. Then, by monotonicity, for any $t \in [T_{(i-1)}, T_{(i)}]$ we have $t \le T_{(i)}$ and hence

$$K_\lambda(t) \;\le\; K_\lambda(T_{(i)}) \;\le\; \ell_{(i)}(\lambda).$$

Integrating against ϕ yields the pathwise inequality

$$\begin{aligned}\rho_\phi(\lambda) &= \int_0^1 K_\lambda(t)\phi(t)\,\mathrm{d}t \\ &\le \sum_{i=1}^{n+1} \ell_{(i)}(\lambda)\big(F_\phi(T_{(i)}) - F_\phi(T_{(i-1)})\big) \\ &= L_\phi^+(\lambda),\end{aligned}$$

which proves the upper bound. For the equality condition, for any fixed $\lambda \in \Lambda$, we define

$$K_\lambda^\star(t) = \sum_{i=1}^{n+1} \ell_{(i)}(\lambda)\, \mathbb{1}_{(T_{(i-1)},T_{(i)}]}(t).$$

Note that this constructed $K_\lambda^\star$ depends on the *same random order statistics* $T_{(i)}$ used to define W_i. With probability one in the draw of the order statistics this function is nondecreasing and matches the calibration losses at the observed ranks. Moreover,

$$\int_0^1 K_\lambda^\star(t)\phi(t)\,\mathrm{d}t = L_\phi^+(\lambda),$$

so $K_\lambda^\star$ is the *least-favorable* quantile function among those compatible with the order constraints (this is the same idea used by CRC/RCPS envelopes [5, 7]). Therefore $L_\phi^+(\lambda)$ is tight: it is the least upper bound compatible with the calibration order constraints. □

B.3. Classical Empirical Quantile Control

Proposition 3.1 (Empirical quantile control)**.** *For any fixed $\lambda \in \Lambda$, let $\widehat{Q}_{1-\delta}(\lambda)$ denote the empirical $(1-\delta)$-quantile of $\{L_{\phi,m}^+(\lambda)\}_{m=1}^M$ obtained in Algorithm 1. For any η with $0 < \eta \le \delta$, the sharp DKW inequality [26] yields*

$$\mathbb{P}\big\{Q_{1-\delta}(\lambda) \le \widehat{Q}_{1-\delta+\eta}(\lambda)\big\} \;\ge\; 1 - 2\exp\big(-2M\eta^2\big), \tag{12}$$

where $\widehat{Q}_{1-\delta+\eta}(\lambda)$ is the empirical quantile at level $(1-\delta+\eta)$. Consequently, choosing $\eta = \sqrt{\frac{\log(2/\gamma)}{2M}}$ guarantees that, for this λ, the underestimation event $Q_{1-\delta}(\lambda) > \widehat{Q}_{1-\delta+\eta}(\lambda)$ occurs with probability at most γ, provided this choice satisfies $\eta \le \delta$.

Proof. Let F_λ be the cumulative distribution function of $L_\phi^+(\lambda)$ and $\widehat{F}_{\lambda,M}$ the empirical distribution function formed by the M Monte Carlo samples. The sharp DKW inequality [26] states that $\mathbb{P}\left[\sup_x |\widehat{F}_{\lambda,M}(x) - F_\lambda(x)| > \eta\right] \le 2e^{-2M\eta^2}$. On the complement event we have $\widehat{F}_{\lambda,M}(Q_{1-\delta}(\lambda)) \ge F_\lambda(Q_{1-\delta}(\lambda)) - \eta \ge 1 - \delta - \eta$, so the empirical $(1-\delta+\eta)$-quantile dominates $Q_{1-\delta}(\lambda)$. Rearranging the DKW bound gives the explicit choice of η. □

B.4. Binomial LCB Guarantee

In this subsection we formalize our binomial LCB acceptance rule.

Idealised vs. implemented Monte Carlo. All guarantees below assume an oracle that produces i.i.d. draws $L_{\phi,1}^+(\lambda), \dots, L_{\phi,M}^+(\lambda) \sim \mathrm{Law}(L_\phi^+(\lambda))$ for each λ. Our implementation approximates this oracle with the proxy $\widetilde{L}_\phi^+(\lambda)$ built from i.i.d. uniform order statistics; this affects only numerical approximation, not the logical form of the guarantees. Fix λ and α, and define the non-exceedance probability

$$q(\lambda;\alpha) := \mathbb{P}\big(L_\phi^+(\lambda) \le \alpha\big).$$

Given M i.i.d. draws from $\mathrm{Law}(L_\phi^+(\lambda))$, form the binomial count

$$S_\lambda(\alpha) = \sum_{m=1}^{M} \mathbb{1}\{L_{\phi,m}^+(\lambda) \le \alpha\} \sim \mathrm{Binomial}\big(M, q(\lambda;\alpha)\big).$$

We compute the Clopper–Pearson lower bound $\underline{q}_\lambda$ at level $(1-\gamma)$. To verify whether $q(\lambda;\alpha) \ge 1-\delta$ (equivalently $Q_{1-\delta}(\lambda) \le \alpha$), our acceptance rule returns "Accept" if $\underline{q}_\lambda \ge 1-\delta$, and returns "Inconclusive" otherwise. We have the following confidence guarantee.

Proposition B.1 (Binomial LCB validity)**.** *If $q(\lambda;\alpha) < 1-\delta$, then the acceptance rule accepts with probability $\le \gamma$. Equivalently, whenever the acceptance rule accepts, it is wrong with probability $\le \gamma$.*

Proof. The Clopper–Pearson lower bound $\underline{q}_\lambda$ at level $(1-\gamma)$ satisfies

$$\mathbb{P}\big(q(\lambda;\alpha) \ge \underline{q}_\lambda\big) \ge 1-\gamma,$$

where $q(\lambda;\alpha) = \mathbb{P}\big(L_\phi^+(\lambda) \le \alpha\big)$. If we actually have $q(\lambda;\alpha) < 1-\delta$, then

$$\mathbb{P}[\text{Accept}] = \mathbb{P}\big(\underline{q}_\lambda \ge 1-\delta\big) \le \mathbb{P}\big(\underline{q}_\lambda > q(\lambda;\alpha)\big) \le \gamma,$$

as desired. □

Note that since $\underline{q}_\lambda$ is a (high-probability) lower bound, we cannot conclude anything if it is less than $1-\delta$, so we return "Inconclusive" in the acceptance rule in that case.

As a final remark, note that $Q_{1-\delta}(\lambda)$ and α are deterministic once λ is fixed; all probabilities in the verification are solely over the Monte Carlo draws used to form $\underline{q}_\lambda$.

B.5. Uniform-over-Grid Monte Carlo Validity

Proposition B.2 (Uniform-over-grid Monte Carlo validity)**.** *Fix α, δ and a finite grid $\mathcal{H} \subset \Lambda$ (or $\mathcal{H} \subset \Lambda \times \Phi$ for multiple spectral densities). For each $h \in \mathcal{H}$, define*

$$q(h;\alpha) := \mathbb{P}\big(L_\phi^+(h) \le \alpha\big),$$
$$Q_{1-\delta}(h) := \inf\{a \in \mathbb{R} : \ \mathbb{P}\big(L_\phi^+(h) \le a\big) \ge 1-\delta\}.$$

Run Algorithm 1 with one-sided binomial LCBs at level $(1-\gamma/|\mathcal{H}|)$ for each $h \in \mathcal{H}$. Then, with probability $\ge 1-\gamma$ over Monte Carlo randomness,

$$\forall h \in \mathcal{H}: \quad \underline{q}_h \ge 1-\delta \Rightarrow Q_{1-\delta}(h) \le \alpha.$$

Hence the data-dependent choice $\widehat{h}$ is valid.

Proof. Proposition B.1 guarantees that, for each fixed $h \in \mathcal{H}$, the event $\{\underline{q}_h \ge 1-\delta \text{ while } q(h;\alpha) < 1-\delta\}$ has probability at most $\gamma/|\mathcal{H}|$ (because the Clopper–Pearson bound for h is computed at level $1-\gamma/|\mathcal{H}|$). A union bound shows that with probability at least $1-\gamma$ over the Monte Carlo randomness, no such event occurs for any $h \in \mathcal{H}$. On that high-probability event, $\underline{q}_h \ge 1-\delta$ implies $q(h;\alpha) \ge 1-\delta$, which is equivalent to $Q_{1-\delta}(h) \le \alpha$ by definition of quantiles. Since $\widehat{h}$ is chosen among the accepted grid points, it inherits the same guarantee. (An alternative approach uses anytime-valid e-processes [32] to provide uniform guarantees without Bonferroni adjustment, useful when $|\mathcal{H}|$ is very large. In the absence of this change, we describe our guarantees as *familywise-valid over a fixed finite grid*.) □

B.6. BQ-SRC Algorithm Correctness

Theorem 3.2 (Calibration-time risk control)**.** *Under assumptions (A1) - (A6), suppose we follow the oracle version of Algorithm 1 and assume additionally that*

$$\{\lambda \in \Lambda_{\text{grid}} : \ \underline{q}_\lambda \ge 1-\delta\} \ne \varnothing, \tag{17}$$

and deploy

$$\widehat{\lambda}_\phi := \min\big\{\lambda \in \Lambda_{\text{grid}} : \ \underline{q}_\lambda \ge 1-\delta\big\}, \tag{18}$$

where $\underline{q}_\lambda$ is the Clopper–Pearson lower bound computed at level $(1-\gamma/|\Lambda_{\text{grid}}|)$. Then, with probability at least $1-\gamma$ over the Monte Carlo randomness used to form the bounds, we can certify that the selected $\widehat{\lambda}_\phi$ satisfies $\rho_\phi(\widehat{\lambda}_\phi) \le \alpha$.

Proof. Fix the calibration sample and associated ordered losses $\{\ell_{(i)}(\lambda)\}_{i=1}^n$. By Theorem 3.1 we have the almost-sure bound

$$\rho_\phi(\lambda) \ \le \ L_\phi^+(\lambda) \quad \text{for every } \lambda \in \Lambda_{\text{grid}}. \tag{21}$$

For $\alpha \in \mathbb{R}$ and $\lambda \in \Lambda_{\text{grid}}$, write

$$q(\lambda;\alpha) := \mathbb{P}\big(L_\phi^+(\lambda) \le \alpha\big),$$

$$S_\lambda(\alpha) := \sum_{m=1}^{M} \mathbb{1}\{L_{\phi,m}^+(\lambda) \le \alpha\},$$

where $L_{\phi,1}^+(\lambda), \dots, L_{\phi,M}^+(\lambda) \overset{\text{i.i.d.}}{\sim} \text{Law}(L_\phi^+(\lambda))$. Then $S_\lambda(\alpha) \sim \text{Binomial}(M, q(\lambda;\alpha))$.

Proposition B.2 (applied with per-grid Clopper–Pearson level $1-\gamma/|\Lambda_{\text{grid}}|$) implies that, with probability at least $1-\gamma$ over the Monte Carlo draws used to form the $\underline{q}_\lambda$'s,

$$\underline{q}_\lambda \ge 1-\delta \quad \Longrightarrow \quad q(\lambda;\alpha) \ge 1-\delta \quad \text{for all } \lambda \in \Lambda_{\text{grid}}.$$

On this high-probability event, the selected $\widehat{\lambda}_\phi$ satisfies

$$\mathbb{P}\big(L_\phi^+(\widehat{\lambda}_\phi) \le \alpha\big) = q(\widehat{\lambda}_\phi;\alpha) \ge 1-\delta \ > \ 0.$$

Combining this with (21) at $\lambda = \widehat{\lambda}_\phi$, Lemma B.4 yields

$$\rho_\phi\big(\widehat{\lambda}_\phi\big) \ \le \ \alpha.$$

The conclusion is deterministic; the Monte Carlo randomness only governs whether the event $q(\widehat{\lambda}_\phi;\alpha) \ge 1-\delta$ holds, and this event fails with probability at most γ. □

C. Batch Multivalid Extension

Let $\mathcal{G}$ be a finite collection of (possibly intersecting) *batch groups* in feature space (e.g., bins of prediction entropy) and $\mathcal{A} \subset [0,1]$ a finite grid of risk targets.

Let X denote the feature component of $Z = (X, Y)$. For each $(\lambda, g, a) \in \Lambda \times \mathcal{G} \times \mathcal{A}$, define the group-conditional non-exceedance probability

$$q_g(\lambda;a) := \mathbb{P}\left[L_{\phi,g}^+(\lambda) \le a \ \middle| \ X \in g\right].$$

Run the same MC test with per-hypothesis level $\gamma/(|\Lambda||\mathcal{G}||\mathcal{A}|)$ and accept (λ, a) as valid for group g if the CP lower bound $\underline{q}_g(\lambda;a) \ge 1-\delta$. Then

$$\mathbb{P}\Big[\exists \lambda, g, a: \ \text{accept but } q_g(\lambda;a) < 1-\delta\Big] \le \gamma,$$

so with probability at least $1-\gamma$, all accepted triples are simultaneously valid.

Remark. This is a finite-grid, calibration-conditional guarantee. For broader multivalid conformal coverage and constructions, see [17].

D. Binomial Intervals Used in Our Tests

Given $S \sim \mathrm{Binom}(M, p)$:

Clopper-Pearson (exact) one-sided LCB at level $1-\gamma$:

$$\mathrm{LCB}_\gamma(p) \;=\; \mathrm{Beta}^{-1}\big(\gamma;\, S,\, M-S+1\big).$$

Wilson score (display only):

Let $z = z_{1-\gamma}$; then the two-sided Wilson interval is

$$\frac{\hat{p} + \frac{z^2}{2M} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{M} + \frac{z^2}{4M^2}}}{1 + \frac{z^2}{M}}, \qquad \hat{p} = \frac{S}{M}.$$

We use Clopper-Pearson in proofs and Wilson for visualization, following recommendations in [8].

E. Experimental Protocol Details

E.1. Implementation details

We implement BQ-SRC and all baselines in a numerically stable, double-precision setting, using vectorized linear algebra for envelope evaluation and Monte Carlo certification. Calibration, Monte Carlo, and test-time evaluation all use the same loss definitions and rescaling (losses are always mapped to $[0, 1]$ before applying spectral risk functionals).

Envelope construction. Given a calibration set of n examples, we sort the calibration losses for each candidate parameter λ to obtain $\ell_{(1)}(\lambda) \le \cdots \le \ell_{(n)}(\lambda)$ and append $\ell_{(n+1)}(\lambda) = 1$. For a given spectral density ϕ with antiderivative $F_\phi(u) = \int_0^u \phi(s)\,\mathrm{d}s$, we idealize sampling from the PIT order statistics $T_{(i)}$ introduced above. In the implementation we approximate these PIT spacings by drawing M i.i.d. uniform samples on $[0, 1]$, sorting them to obtain order statistics $T_{(1)}, \ldots, T_{(n)}$, and defining spacings

$$W_i \;=\; F_\phi\big(T_{(i)}\big) - F_\phi\big(T_{(i-1)}\big), \qquad T_{(0)} := 0.$$

The upper envelope is then

$$L^+_\phi(\lambda) \;=\; \sum_{i=1}^{n+1} W_i\, \ell_{(i)}(\lambda),$$

and the calibration-conditional guarantee in Theorem 3.1 is enforced exactly in this discretized form. We reuse the same set of order statistics $T_{(1:n)}$ across grid points λ in a given experiment to reduce Monte Carlo variance, but $T_{(1:n)}$ is always independent of the calibration data.

Spectral densities. All spectra satisfy $\phi \ge 0$, are non-decreasing, and integrate to 1:

- **Mean risk**: $\phi(t) \equiv 1$.
- **CVaR**$_\beta$ ($\beta \in [0, 1)$): $\phi(t) = \frac{1}{1-\beta}\,\mathbf{1}\{t > \beta\}$.
- **Mixtures**: for mix 95/5 we use

$$\phi_{\text{mix95/5}}(t) = 0.95 \cdot 1 \;+\; 0.05 \cdot \frac{1}{1-0.9}\,\mathbf{1}\{t > 0.9\},$$

  and analogously for mix 90/10 and mix 85/15 (replacing the outer weight and CVaR weight accordingly). These spectra put most mass on the mean risk while allocating a small fraction of mass to a $\mathrm{CVaR}_{0.9}$-type tail.
- **Ramped tails**: for ramped-tail spectra, we use a density that is flat on $[0, \beta]$ and increases as a power function on $(\beta, 1]$, with parameters (β, κ) (e.g., $\beta = 0.7$, $\kappa = 2.0$) chosen so that $\int_0^1 \phi = 1$ and ϕ is non-decreasing.

These closed-form densities are used consistently across all experiments and in the proofs in Appendix A.1.

Monte Carlo certification and binomial LCBs. For each candidate λ and target risk threshold α, we approximate the non-exceedance probability

$$q(\lambda; \alpha) \;=\; \Pr\big\{L^+_\phi(\lambda) \le \alpha\big\}$$

using M independent Monte Carlo draws of the envelope. In all main experiments we set

$$M = 5000, \qquad \gamma = 10^{-3}, \qquad 1 - \delta = 0.95,$$

and we consider ablations with $M \in \{2000, 5000, 10000\}$ and $\delta \in \{0.05, 0.1\}$ (see Figure 3 and Figure 7). For each (λ, α) we form a success count

$$S_\lambda(\alpha) = \sum_{m=1}^{M} \mathbf{1}\big\{L^+_{\phi,m}(\lambda) \le \alpha\big\} \sim \mathrm{Binomial}\big(M, q(\lambda; \alpha)\big),$$

and compute the one-sided $(1 - \gamma_{\text{per-test}})$ Clopper–Pearson lower confidence bound

$$\underline{q}_\lambda = \mathrm{Beta}^{-1}\big(\gamma_{\text{per-test}}; S_\lambda(\alpha), M - S_\lambda(\alpha) + 1\big).$$

We accept λ if $\underline{q}_\lambda \ge 1 - \delta$, which is equivalent to certifying that the $(1-\delta)$-quantile $Q_{1-\delta}(\lambda)$ of $L^+_\phi(\lambda)$ is at most α with confidence at least $1 - \gamma_{\text{per-test}}$.

To obtain uniform control over finite grids $\mathcal{H}$ (e.g., Λ_{grid} or a two-parameter grid $\mathcal{G}$), we apply a Bonferroni-adjusted level

$$\gamma_{\text{per-test}} = \gamma / |\mathcal{H}|$$

and run the same binomial LCB test for each grid point in $\mathcal{H}$. Proposition B.2 shows that, with probability at least $1 - \gamma$, all accepted grid points simultaneously satisfy $Q_{1-\delta}(\lambda) \le \alpha$. In the batch multivalid extension, δ is deliberately kept global rather than Bonferroni-split across groups; validity is obtained by a finite union bound over discretized exceedance thresholds instead (see the discussion under Batch Multivalid Calibration).

Grid search and two-parameter control. For single-parameter control, we discretize the λ-axis into $|\Lambda_{\text{grid}}| \in \{128, 256, 512\}$ evenly spaced thresholds on an interval $[\lambda_{\min}, \lambda_{\max}]$ with $\lambda_{\min}$ typically in $\{0.30, 0.35\}$ for COCO and chosen to cover the relevant operating range in other tasks (e.g., from 0 up to a maximum score or interval width). For two-parameter (τ, k) control we form a rectangular grid $\mathcal{G}$ with

$$\tau \in \{0, \Delta_\tau, 2\Delta_\tau, \ldots, \tau_{\max}\}, \quad k \in \{1, 2, \ldots, k_{\max}\},$$

with τ-grid size 64 and $k_{\max} = 15$ in the COCO experiments. For each grid point we evaluate the envelope, run the binomial LCB test described above, and then select among the accepted grid points the one with smallest empirical efficiency metric (interval length, prediction-set size, or analogous quantity), consistent with the monotonicity properties in Proposition B.3.

Randomization and shared randomness. Across all experiments we fix a base random seed for each trial, and we reuse the same calibration/test split and Monte Carlo order statistics across methods within that trial. This ensures that differences between methods are not driven by differences in the underlying randomness. All reported means and confidence intervals in the tables are empirical averages and binomial or t-intervals computed over the stated number of independent trials (e.g., 10,000 trials for synthetic $n = 10$, 2,000 for $n = 200$, 1,000 for heteroskedastic and COCO, and 10 for segmentation).

E.2. MS-COCO Protocol

For MS-COCO multilabel classification experiments:

- **Dataset**: MS-COCO 2017 splits [23] with standard preprocessing
- **Base model**: Following [39], We adopt the logits of a ResNet-based architecture with sigmoid head outputting scores over 80 classes
- **Training**: Standard ImageNet pretraining with COCO fine-tuning (standard recipe)
- **Calibration/test sizes**: Calibration set size $n = 1000$; test set uses standard COCO splits
- **Common hyperparameters**: $\alpha = 0.1$ (sweeps use $\alpha \in \{0.05, 0.1, 0.2\}$), $1 - \delta = 0.95$ ($\delta = 0.05$, sweeps use $\delta \in \{0.05, 0.1\}$), $M \in \{2000, 5000, 10000\}$ Monte Carlo draws (main results use $M = 5000$), $\gamma = 10^{-3}$
- **Quantile method**: Binomial LCB (Clopper-Pearson or Wilson) [10, 43]
- **Sampling**: Antithetic sampling mode for variance reduction
- **Temperature scaling**: Applied when specified (see temperature ablation experiments); uses calibration set with step size 0.05 and max iterations 200 when explicitly configured [15]
- **Grid sizes**:
  - Single-parameter: $|\Lambda_{\text{grid}}| \in \{128, 256, 512\}$ with minimum threshold $\lambda_{\min} \in \{0.30, 0.35\}$ (main results use $\lambda_{\min} = 0.30$, $|\Lambda_{\text{grid}}| = 256$)
  - Two-parameter (τ, k): τ grid size 64 ($\tau \in [0, 0.9]$ or $[0, 1]$ depending on experiment), $k \in \{1, \ldots, 15\}$ (step size 1, $k_{\min} = 1$ when specified)
- **Spectral densities** ϕ: Uniform (mean), $\text{CVaR}_{0.9}$, $\text{CVaR}_{0.95}$, mix 95/5 (95% mean risk + 5% $\text{CVaR}_{0.9}$), mix 90/10, mix 85/15, and ramped tail variants (ramp $\beta = 0.7$, $\kappa = 2.0$)
- **Batch multivalid**: Number of strata $|\mathcal{G}| \in \{8, 16\}$; stratification by prediction entropy or confidence score
- **Shift-aware weighting**: Uses inverse-positive-count importance weights for covariate shift experiments
- **Number of trials**: 1000 trials per experiment configuration
- **Seeds**: Random seeds fixed across methods for fair comparison; code repository will be made available

E.3. Synthetic benchmark protocol

For the synthetic spectral risk-control experiments:

- **Loss model**: We draw $U \sim \text{Unif}(0, 1)$ and define a scaled count loss $\ell(z; \lambda) = \mathbf{1}\{U > \lambda\}$, which induces a simple monotone loss in the control parameter $\lambda \in [0, 1]$.
- **Calibration and test sizes**: We consider calibration sizes $n \in \{10, 200\}$ and, for each n, generate 10,000 (for $n = 10$) or 2,000 (for $n = 200$) independent calibration/test splits. In each split, n examples are used for calibration and the remaining 10,000 (for $n = 10$) or 5,000 (for $n = 200$) for evaluation.
- **Target risk**: The target spectral risk is fixed at $\alpha = 0.4$ with confidence level $1 - \delta = 0.95$ in all synthetic runs.
- **Methods and spectra**: We evaluate CRC, RCPS, HPD, and several BQ-SRC variants: spectral mean (uniform ϕ), $\text{CVaR}_{0.9}$, and mixture spectra such as mix 95/5 (95% mean risk + 5% $\text{CVaR}_{0.9}$). Results are summarized in Tables 1–2.
- **Evaluation**: For each method and spectrum, we compute the empirical spectral risk and violation rate over all test examples and trials, and report the mean and 95% confidence intervals across trials. The decision parameter λ is chosen from a uniform grid on $[0, 1]$ using the binomial LCB rule described above.

E.4. Heteroskedastic regression protocol

For heteroskedastic regression:

- **Data generation**: Drawing covariates $X \sim \text{Unif}([0, 4])$ and responses $Y \sim \mathcal{N}(0, |X|)$, and then working with $|Y|$ as the relevant magnitude. Calibration uses $n = 200$ examples; the remaining examples serve as test points.
- **Prediction intervals and loss**: For each trial we construct symmetric prediction intervals centered at 0 with

half-width λ; the deployment loss is the miscoverage indicator of the interval. Efficiency is measured by the average interval length 2λ on the test set.

- **Target risk and trials**: The target risk is $\alpha = 0.1$ with confidence level $1 - \delta = 0.95$. We run 1,000 random calibration/test splits per method.
- **Methods and spectra**: We compare CRC, RCPS, HPD, and BQ-SRC (primarily the mix 95/5 spectral density) under the same calibration data. The main summary appears in Table 3.
- **Calibration grid**: We discretize the half-width λ over a uniform grid spanning the empirically relevant scale of the response (from a small positive minimum up to a large maximum ensuring near-trivial coverage), and select the smallest λ whose certified spectral risk does not exceed α.

E.5. Closed-set ImageNet protocol

For closed-set classification on ImageNet:

- **Dataset and model**: We use the ImageNet validation split with a ResNet-152 classifier as the base model. The classifier is pre-trained on the ImageNet training set using a standard training recipe.
- **Nonconformity scores**: We evaluate three standard scores: LAC, APS, and RAPS [3], following prior work on risk-controlling prediction sets.
- **Calibration sizes and target risks**: Calibration sizes are $n \in \{256, 1000, 5000\}$; target risks $\alpha \in \{0.1, 0.05, 0.01\}$ with confidence level $1 - \delta = 0.95$.
- **Methods**: We include Split-CP APS baselines, CRC, RCPS, and BQ-SRC with several spectral densities (mean, mix 95/5, and other mixtures), along with ablations that swap binomial LCB quantiles for DKW inflation. For each method, we evaluate coverage, mean set size, and class-conditional coverage (CCV) variance at each (α, n) combination.
- **Evaluation**: For each configuration we form prediction sets using the chosen nonconformity score and calibration rule, then compute empirical coverage and set size over the full validation set. Closed-set tables in Section G report averages across all runs at each (α, n) and include CCV variance as a fairness-sensitive metric.

E.6. Zero-shot classification protocol

For zero-shot experiments:

- **Backbones and adaptations**: We consider two CLIP-family encoders, OpenAI CLIP ViT-L/14 and MetaCLIP ViT-H/14, following framework of [38].
- **Datasets**: We adopt the standard CLIP-Conformal zero-shot suites, grouped into: ImageNet and its variants (ImageNet, V2, R, Sketch), shifted/robustness benchmarks (ImageNet-A, SUN397, Aircraft, EuroSAT, Stanford Cars), and long-tail datasets (Food101, Oxford Pets, Flowers, Caltech-101, DTD, UCF).
- **Calibration protocol**: For each dataset/backbone/adaptation combination, we consider target risks $\alpha \in \{0.1, 0.05\}$, confidence level $1 - \delta = 0.95$, and calibration fractions in $\{1\%, 2.5\%, 5\%, 10\%\}$, using 20 random seeds. APS is used as the nonconformity score for Split-CP baselines and BQ-SRC.
- **Spectral densities and risk multipliers**: BQ-SRC is run with spectral mean, mix 95/05, and $\mathrm{CVaR}_{0.9}$ densities and with risk multipliers $\{1.0, 0.75, 0.5\}$ scaling the target risk.
- **Evaluation and aggregation**: For each configuration we compute per-dataset coverage, mean prediction-set size, and tail metrics such as $\mathrm{CVaR}_{0.95}$ of the miscoverage distribution. The zero-shot tables in this appendix aggregate these metrics by averaging over calibration fractions and seeds, reporting one row per dataset/backbone/adaptation/method combination.

E.7. Semantic segmentation protocol

For semantic segmentation experiments:

- **Datasets and models**: We evaluate Cityscapes with PSPNet logits and ADE20K with SegFormer logits. The base models are trained on the standard training splits with widely used configurations; segmentation logits and ground-truth masks are then treated as fixed inputs to calibration.
- **Losses and masks**: We consider two loss types: (i) a binary minimum-coverage loss that enforces per-image coverage constraints, and (ii) a pixel-wise miscoverage loss, following the definitions in [27].
- **Calibration sizes and trials**: For ADE20K we use calibration size $n = 512$ and 10 randomized calibration/test splits; Cityscapes experiments use calibration size $n = 256$ with the same number of splits. In each trial, distinct calibration and validation subsets are drawn without replacement from the corresponding validation set.
- **Hyperparameters**: For binary-loss experiments, we sweep minimum coverage ratios (e.g., 0.75 and 0.90 on ADE20K, 0.95 and 0.99 on Cityscapes) and miscoverage targets α in $\{0.01, 0.05, 0.10, 0.20\}$ for ADE20K and a similar range (including $\alpha \in \{0.005, 0.01, 0.05\}$) for Cityscapes. Spectral densities include mean, mix 95/5, and $\mathrm{CVaR}_{0.9}$.
- **Evaluation**: For each configuration (dataset, loss type, multimask, α, coverage target, spectrum) we form calibrated multimask predictions and compute empirical risk, coverage, and mean set size per pixel, averaging across images and the 10 trials. ADE20K tables in this appendix report these metrics, while Cityscapes and LoveDA summary statistics are collected in the final segmentation table.

F. Additional Experimental Results

Risk–Efficiency Trade-off. Fig. 2 compares prediction-set size and spectral risk across methods on MS-COCO under the specified parameters, showing that BQ-SRC-LCB significantly outperforms RCPS while closely tracking the target error level.

Temperature scaling ablation. Figure 8 compares BQ-SRC with and without temperature scaling [15]. Temperature scaling reduces mean set size by ≈ 5–10% while maintaining validity, demonstrating that better-calibrated scores improve efficiency without weakening guarantees.

Grid size and MC budget robustness. Figure 5 shows violation rates and set sizes across grid sizes $\{128, 256, 512\}$ and MC budgets $\{2{,}000, 5{,}000, 10{,}000\}$. Violation rates remain below $\delta = 5\%$ across all configurations when γ is properly adjusted by the union bound ($\gamma_{\text{per-test}} = \gamma/|\Lambda_{\text{grid}}|$), demonstrating robustness to hyperparameter choices.

Batch multivalid BQ-SRC. Figure 6 compares global BQ-SRC to batch multivalid BQ-SRC with 8 bins stratified by prediction entropy. The entropy model drives the empirical violation rate to **0.0%** (vs. $\delta = 5\%$) while adapting thresholds across difficulty bins, at the cost of a modest increase in mean set size (3.86 vs. 3.39). Using confidence-based strata yields slightly smaller sets (3.44–3.45) with 1.4–2.6% violations, still well below the nominal δ budget. Importantly, δ is *not split* across bins via Bonferroni; instead, validity follows from a finite union over discretized exceedance thresholds coupled with the LCB test, so δ remains global [17]. These results show that the multivalid extension enforces subgroup guarantees without resorting to Bonferroni splits of δ.

Parameter sweeps. Figure 7 shows set sizes and violation rates across $\alpha \in \{0.05, 0.1, 0.2\}$ and $\delta \in \{0.05, 0.1\}$. Results demonstrate monotonic behavior: larger α and δ permit smaller sets, while violations remain controlled at or below the nominal δ level.

Tail-shifted synthetic stress test. We additionally probe a tail-shifted synthetic setting in which the calibration and test distributions differ in the right tail of the loss, following the construction used in our internal diagnostics. Table 7 compares CRC and BQ-SRC (mix 95/5) on this experiment, showing that BQ-SRC drives spectral risk essentially to zero while maintaining strong control of tail events.

F.1. Batch Multivalid Calibration

F.2. Shift-Aware Calibration

Proposition F.1 (Weighted vs unweighted violations)**.** *When using importance weights w_i for covariate shift, the guarantee holds for the* weighted risk $\mathbb{E}_w[\ell(Z;\lambda)] = \sum_i w_i \ell(Z_i;\lambda)$ *where* $\sum_i w_i = 1$. *The* unweighted violation rate $\mathbb{P}\left[\rho_\phi(\widehat{\lambda}) > \alpha\right]$ *is not controlled and can exceed δ. A worst-case bound: if weights are bounded away from zero ($w_i \geq w_{\min} > 0$) and the unweighted empirical risk is $R = \frac{1}{n}\sum_i \ell_i$, then the weighted risk satisfies*

$$R_w = \sum_i w_i \ell_i \;\geq\; w_{\min} \sum_i \ell_i \;=\; n w_{\min} R,$$

but this does not provide a useful upper bound on unweighted violations. A simple counterexample: with two groups where group 1 has high loss and weight $w_1 \approx 1$, group 2 has low loss and weight $w_2 \approx 0$, the weighted risk can be controlled while the unweighted violation rate approaches 1.

G. Closed-Set ImageNet Tables

Tables 10–18 summarize every configuration in our closed-set ImageNet study, covering three nonconformity scores (LAC, APS, RAPS), three calibration sizes ($n \in \{256, 1000, 5000\}$), and multiple spectral/ablation variants for each method. For each (α, n) and method, we report mean coverage, mean prediction-set size, and class-conditional coverage (CCV) variance over repeated trials; these metrics are computed from the full ImageNet validation set under the calibrated prediction sets.

H. Zero-Shot Tables

Aggregation protocol. Tables 19–21 summarize every zero-shot experiment in our study. For each dataset/backbone/adaptation configuration we run Split-CP and BQ-SRC at both target risks ($\alpha \in \{0.1, 0.05\}$), across calibration fractions in $[1\%, 10\%]$, and over 20 random seeds. We then aggregate coverage, mean set size, and $\text{CVaR}_{0.95}$ across α, calibration fractions, and seeds, reporting one row per dataset/backbone/adaptation/method combination. Each row lists the Split-CP APS baseline plus three BQ-SRC variants (spectral mean, mix95/05, and $\text{CVaR}_{0.9}$, all at risk multiplier 1).

Key trends. On ImageNet-style robustness suites (Table 19) BQ-SRC drives coverage from roughly 0.93 to 0.996 independent of the spectral choice, while Conf-OT primarily benefits CLIP-ViT-L/14 by slashing tail risk (e.g., imagenet-sketch $\text{CVaR}_{0.95}$: $0.101 \rightarrow 0.004$ for the

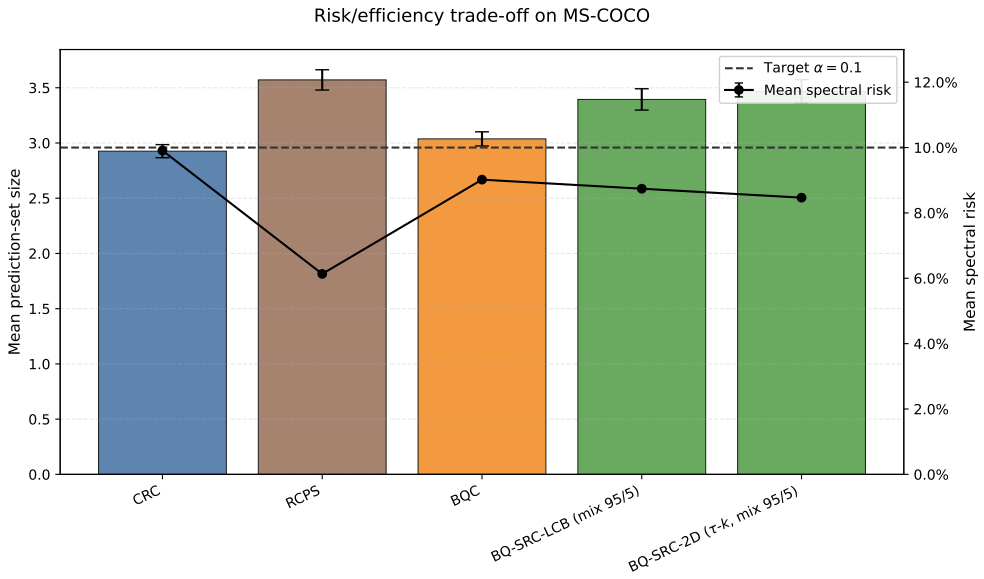

Figure 2. MS-COCO trade-offs at $\alpha = 0.1$, $1 - \delta = 0.95$, $M = 5000$, $\gamma = 10^{-3}$. Bars: mean prediction-set size $\pm$ one standard deviation. Line: mean spectral risk with 95% Wilson confidence intervals; dashed line denotes target $\alpha = 0.1$. Legend: BQC.

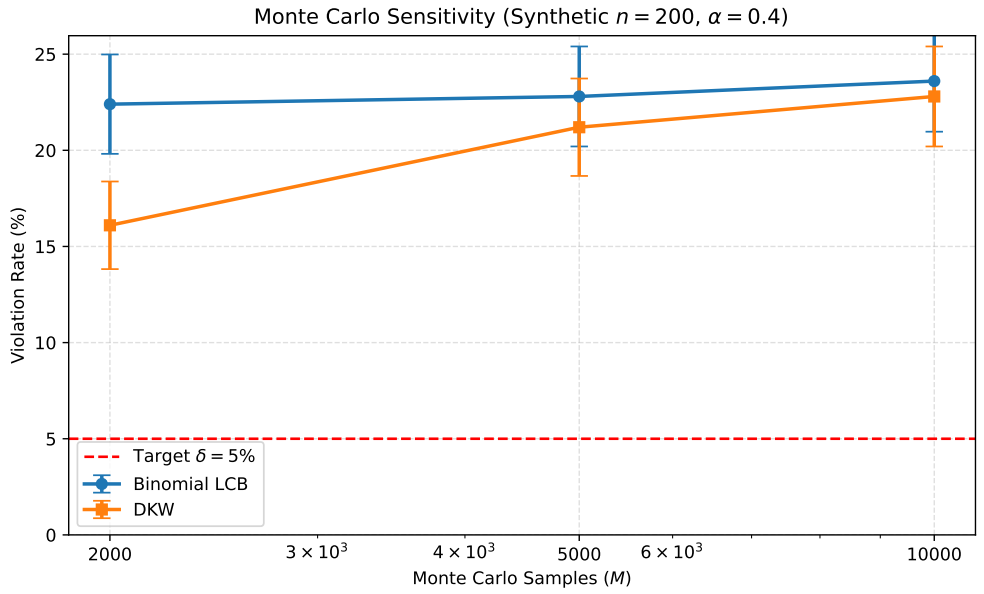

Figure 3. Monte Carlo sensitivity on synthetic $n = 200$ benchmark ($\gamma = 10^{-3}$, grid size $|\Lambda_{\text{grid}}| = 256$, $\delta = 0.05$). Shaded regions denote 95% Wilson confidence intervals [43]; dashed line marks target $\alpha = 0.4$. Common random numbers are reused across grid points. The binomial LCB method (using Clopper–Pearson [10] or Wilson [43]) provides tighter control than DKW inflation across all budgets.

$\text{CVaR}_{0.9}$ spectral). The shifted benchmarks (Table 20) exhibit the largest Conf-OT deltas—ImageNet-A $\text{CVaR}_{0.95}$ falls from 0.11 to 0.002 even for the smoother spectral mixtures. Long-tail datasets (Table 21) show smaller gains; MetaCLIP without Conf-OT already delivers > 0.99 coverage, so the adaptation can be skipped when efficiency is paramount and spectral mean/mix suffice.

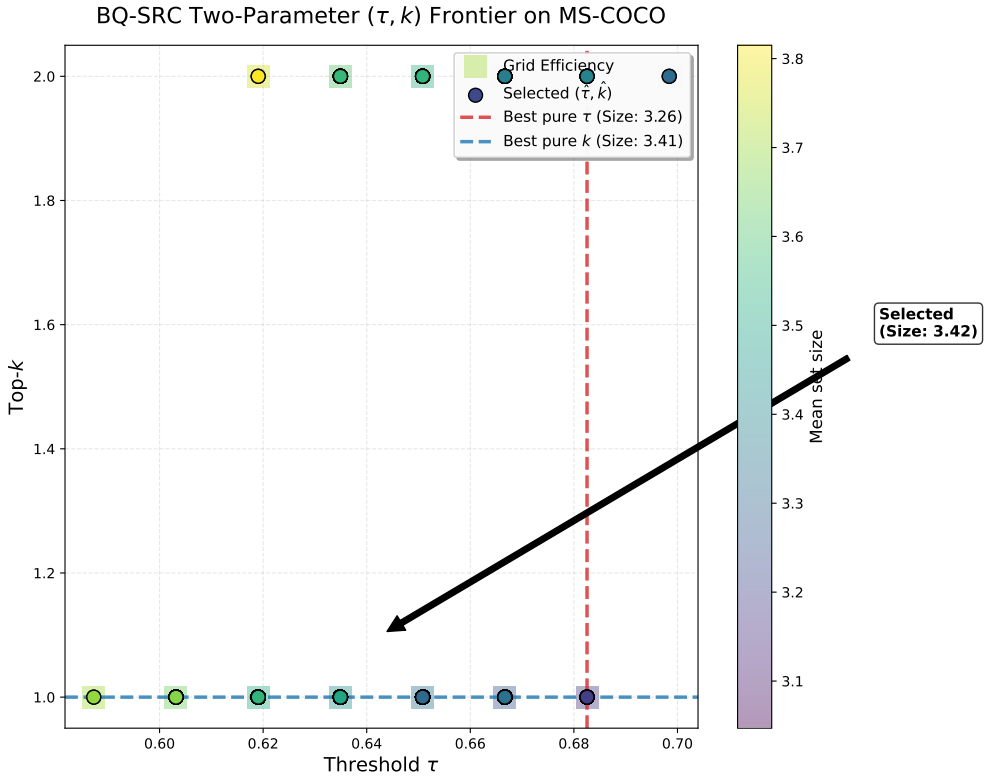

Figure 4. Two-parameter (τ, k) frontier on MS-COCO ($\alpha = 0.1$, $\delta = 0.05$, $M = 5000$, $\gamma = 10^{-3}$). The plot visualizes the efficiency landscape (background heatmap) and the selected $(\hat{\tau}, \hat{k})$ points (foreground scatter) across 1000 trials. Reference lines indicate the best single-parameter optima (pure τ and pure k), annotated with their respective mean prediction-set sizes. The joint control strategy selects points off the axes, achieving a lower mean set size (annotated centroid) and demonstrating dominance over single-parameter controls. The feasible frontier is an up-set under the partial order $(\tau_1, k_1) \preceq (\tau_2, k_2)$ iff $\tau_1 \geq \tau_2$ and $k_1 \leq k_2$ (Proposition B.3). Grid size: $|\mathcal{G}| = 256$ pairs. The k-axis range is motivated by the average COCO label count ≈ 2.9 per image [23].

Table 7. Tail-shifted synthetic experiment (5,000 trials, target spectral risk $\alpha = 0.4$). BQ-SRC with a mix 95/5 spectral density achieves near-zero spectral risk and zero empirical violations, while CRC remains slightly conservative in the tail.

Method	Spectral risk	Violations	λ
CRC (Angelopoulos et al., 2024)	0.011 (0.011, 0.011)	0.00% (0.00, 0.18)	0.597 (0.596, 0.597)
BQ-SRC (mix 95/5)	0.000 (0.000, 0.000)	0.00% (0.00, 0.07)	0.861 (0.860, 0.863)

Table 8. Batch multivalid BQ-SRC on MS-COCO (1,000 trials, $\alpha = 0.1$, $1 - \delta = 0.95$). Spectral risk and prediction-set size are reported as mean (95% CI). Violation rates use 95% Wilson confidence intervals [43]. A single global δ is used (not Bonferroni-split); groups are stratified by prediction entropy.

Stratification	Bins	Spectral risk	Violations	Prediction-set size
Entropy	8	0.0830 (0.0827, 0.0833)	0.00% (0.00, 0.00)	3.855 (3.846, 3.864)
Entropy	16	0.0842 (0.0839, 0.0845)	0.00% (0.00, 0.00)	3.887 (3.877, 3.896)
Confidence	8	0.0879 (0.0876, 0.0883)	1.40% (0.67, 2.13)	3.443 (3.437, 3.449)
Confidence	16	0.0891 (0.0888, 0.0895)	2.60% (1.61, 3.59)	3.455 (3.448, 3.461)

Table 9. Shift-aware BQ-SRC on MS-COCO with inverse-positive-count importance weights (1,000 trials, $\alpha = 0.1$, $1 - \delta = 0.95$). Weighted metrics evaluate the guarantee under the deployed importance weights. Violation rates use 95% Wilson confidence intervals [43]. *Note:* Validity here is with respect to the *weighted* target distribution (via importance sampling); unweighted violations are not controlled (see Proposition F.1). This follows the weighted conformal prediction framework [40] for covariate shift. The validity guarantee is with respect to **weighted risk** under the estimated density ratio; unweighted rates can exceed δ because guarantees are weighted under covariate shift [40]. In our experiments, inverse-positive-count weights are proportional to the reciprocal of the number of positive labels for each example, normalized so that $\sum_i w_i = 1$.

Metric	Unweighted	Weighted
Spectral risk	0.1082 (0.1077, 0.1086)	0.0876 (0.0872, 0.0880)
Violation rate	85.60% (83.42, 87.78)	2.80% (1.78, 3.82)
Prediction-set size	2.436 (2.432, 2.440)	2.436 (2.432, 2.440)

Table 10. Closed-set classification - Adaptive Prediction Sets (APS) [34]: mean coverage across α and calibration sizes.

Alpha			0.010000			0.050000			0.100000
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	1.000	1.000	1.000	0.969	0.962	0.960	0.935	0.930	0.929
BQ-SRC (Mix 95/5)	1.000	1.000	1.000	0.985	0.975	0.973	0.955	0.949	0.947
BQ-SRC (Mix 95/5) + DKW	1.000	1.000	1.000	0.989	0.975	0.973	0.958	0.950	0.948
CRC	0.992	0.990	0.990	0.961	0.958	0.958	0.929	0.927	0.927
RCPS	0.999	1.000	1.000	0.999	0.988	0.971	0.975	0.950	0.937
Split CP	0.993	0.990	0.990	0.961	0.958	0.958	0.929	0.927	0.927

Table 11. Closed-set classification - APS: mean prediction-set size.

Alpha			0.010000			0.050000			0.100000
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	990.469	999.047	1000.000	43.668	17.622	15.994	8.022	6.881	6.601
BQ-SRC (Mix 95/5)	990.469	1000.000	1000.000	318.641	31.049	26.481	14.429	11.236	10.559
BQ-SRC (Mix 95/5) + DKW	990.469	1000.000	1000.000	488.375	33.411	26.787	16.026	11.460	10.603
CRC	130.834	84.340	79.032	17.375	14.839	14.567	6.859	6.424	6.361
RCPS	882.681	1000.000	1000.000	882.681	71.157	23.739	32.984	11.463	7.985
Split CP	141.312	86.500	79.563	17.489	14.812	14.559	6.855	6.411	6.359

Table 12. Closed-set classification - class-conditional coverage (CCV) variance.

Alpha			0.010000			0.050000			0.100000
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	0.000	0.000	0.000	0.001	0.002	0.002	0.004	0.004	0.004
BQ-SRC (Mix 95/5)	0.000	0.000	0.000	0.001	0.001	0.001	0.002	0.003	0.003
BQ-SRC (Mix 95/5) + DKW	0.000	0.000	0.000	0.000	0.001	0.001	0.002	0.002	0.003
CRC	0.000	0.000	0.000	0.002	0.002	0.002	0.004	0.004	0.004
RCPS	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.002	0.003
Split CP	0.000	0.000	0.000	0.002	0.002	0.002	0.004	0.004	0.004

Table 13. Closed-set classification - Least Ambiguous Set-Valued Classifier (LAC) [35]: coverage.

Alpha		0.010000			0.050000			0.100000	
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	1.000	1.000	1.000	0.980	0.961	0.958	0.913	0.904	0.902
BQ-SRC (Mix 95/5)	1.000	1.000	1.000	0.997	0.995	0.998	0.949	0.937	0.935
Split CP	0.992	0.990	0.990	0.953	0.950	0.950	0.903	0.900	0.900

Table 14. Closed-set classification - LAC: prediction-set size.

Alpha		0.010000			0.050000			0.100000	
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	1000.000	1000.000	1000.000	540.025	98.058	4.547	2.147	1.866	1.813
BQ-SRC (Mix 95/5)	1000.000	1000.000	1000.000	923.362	859.696	952.237	89.247	2.894	2.712
Split CP	63.035	28.602	25.825	4.357	3.681	3.615	1.889	1.790	1.776

Table 15. Closed-set classification - LAC: CCV variance.

Alpha		0.010000			0.050000			0.100000	
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	0.000	0.000	0.000	0.001	0.002	0.002	0.006	0.007	0.007
BQ-SRC (Mix 95/5)	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.004	0.004
Split CP	0.000	0.000	0.000	0.003	0.003	0.003	0.007	0.007	0.007

Table 16. Closed-set classification - Regularized APS (RAPS) [3]: coverage.

Alpha		0.010000			0.050000			0.100000	
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	0.996	0.994	0.991	0.969	0.962	0.959	0.934	0.929	0.928
BQ-SRC (Mix 95/5)	0.996	0.996	0.994	0.982	0.975	0.976	0.954	0.948	0.946
Split CP	0.992	0.990	0.990	0.958	0.955	0.955	0.927	0.925	0.925

Table 17. Closed-set classification - RAPS: prediction-set size.

Alpha		0.010000			0.050000			0.100000	
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	177.696	52.229	31.141	9.065	7.591	7.283	5.091	4.704	4.603
BQ-SRC (Mix 95/5)	177.696	102.530	50.651	14.717	9.763	9.690	6.871	6.154	5.991
Split CP	69.366	29.256	25.176	7.249	6.855	6.837	4.587	4.442	4.432

Table 18. Closed-set classification - RAPS: CCV variance.

Alpha			0.010000			0.050000			0.100000
Calibration size	256	1000	5000	256	1000	5000	256	1000	5000
Method									
BQ-SRC (Mean)	0.000	0.000	0.000	0.001	0.002	0.002	0.004	0.004	0.004
BQ-SRC (Mix 95/5)	0.000	0.000	0.000	0.001	0.001	0.001	0.002	0.003	0.003
Split CP	0.000	0.000	0.000	0.002	0.002	0.002	0.004	0.004	0.005

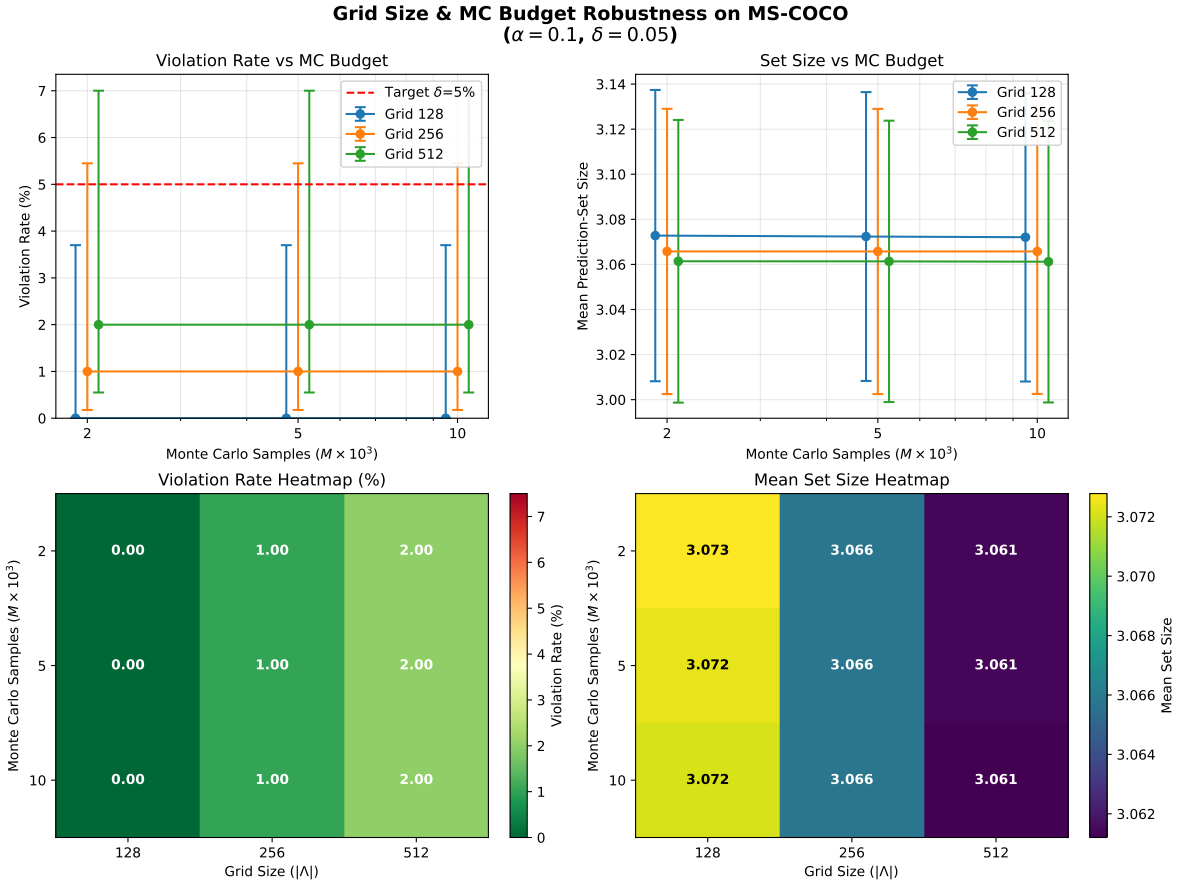

Figure 5. Grid size and Monte Carlo budget robustness on MS-COCO ($\alpha = 0.1$, $\delta = 0.05$, $\gamma = 10^{-3}$). Top: violation rates and set sizes vs MC budget for different grid sizes. Bottom: heatmaps showing stability across configurations. Violation rates remain below $\delta = 5\%$ when γ is adjusted by union bound ($\gamma_{\text{per-test}} = \gamma/|\Lambda_{\text{grid}}|$).

Table 19. Zero-shot accuracy and tail risk on ImageNet, V2, -R, and Sketch averaged over all calibration fractions and $\alpha \in \{0.1, 0.05\}$.

Dataset	Backbone	Adapt	Split CP (APS)			BQ mean			BQ mix95/05			BQ $\mathrm{CVaR}_{0.9}$		
			Cov	Size	$\mathrm{CVaR}_{0.95}$	Cov	Size	CVaR	Cov	Size	CVaR	Cov	Size	CVaR
imagenet	CLIP-ViT-L/14	none	0.928	4.8	0.960	0.944	6.2	0.894	0.958	8.8	0.772	0.995	101.0	0.102
imagenet	CLIP-ViT-L/14	confot	0.928	5.3	0.053	0.943	6.7	0.040	0.958	10.3	0.029	0.994	125.8	0.003
imagenet	MetaCLIP-ViT-H/14	none	0.925	2.9	0.071	0.945	4.0	0.052	0.960	5.6	0.038	0.995	74.6	0.004
imagenet	MetaCLIP-ViT-H/14	confot	0.926	3.1	0.056	0.947	4.4	0.034	0.961	6.1	0.022	0.995	110.5	0.002
imagenetv2	CLIP-ViT-L/14	none	0.924	5.6	0.964	0.938	7.1	0.928	0.953	10.9	0.816	0.995	200.9	0.101
imagenetv2	CLIP-ViT-L/14	confot	0.921	6.2	0.065	0.936	7.8	0.053	0.952	12.5	0.039	0.995	241.5	0.004
imagenetv2	MetaCLIP-ViT-H/14	none	0.925	3.5	0.080	0.941	4.6	0.063	0.957	6.8	0.047	0.995	131.1	0.005
imagenetv2	MetaCLIP-ViT-H/14	confot	0.922	4.1	0.078	0.941	5.6	0.056	0.956	8.3	0.039	0.995	156.9	0.001
imagenet-r	CLIP-ViT-L/14	none	0.931	1.8	0.956	0.941	2.3	0.917	0.956	3.6	0.788	0.995	55.3	0.107
imagenet-r	CLIP-ViT-L/14	confot	0.927	2.0	0.078	0.937	2.5	0.069	0.956	4.3	0.051	0.994	61.8	0.005
imagenet-r	MetaCLIP-ViT-H/14	none	0.951	1.2	0.038	0.958	1.4	0.034	0.965	1.9	0.030	0.995	36.2	0.006
imagenet-r	MetaCLIP-ViT-H/14	confot	0.949	1.2	0.040	0.956	1.4	0.036	0.963	2.0	0.032	0.995	41.9	0.005
imagenet-sketch	CLIP-ViT-L/14	none	0.923	24.6	0.982	0.931	28.2	0.969	0.950	47.0	0.839	0.995	623.2	0.101
imagenet-sketch	CLIP-ViT-L/14	confot	0.921	19.1	0.064	0.929	22.0	0.058	0.950	38.9	0.041	0.994	551.1	0.004
imagenet-sketch	MetaCLIP-ViT-H/14	none	0.924	8.8	0.074	0.936	11.0	0.061	0.954	19.9	0.047	0.994	522.7	0.008
imagenet-sketch	MetaCLIP-ViT-H/14	confot	0.923	8.5	0.074	0.937	11.0	0.060	0.953	18.8	0.048	0.994	459.8	0.005

Table 20. Zero-shot robustness on ImageNet-A, SUN397, Aircraft, EuroSAT, and Stanford Cars.

Dataset	Backbone	Adapt	Split CP (APS)			BQ mean			BQ mix95/05			BQ $\mathrm{CVaR}_{0.9}$		
			Cov	Size	$\mathrm{CVaR}_{0.95}$	Cov	Size	CVaR	Cov	Size	CVaR	Cov	Size	CVaR
imagenet-a	CLIP-ViT-L/14	none	0.935	10.1	0.889	0.945	13.5	0.823	0.965	23.2	0.644	0.997	115.9	0.054
imagenet-a	CLIP-ViT-L/14	confot	0.923	11.8	0.103	0.934	15.4	0.087	0.956	25.3	0.052	0.996	124.3	0.002
imagenet-a	MetaCLIP-ViT-H/14	none	0.933	8.1	0.074	0.942	11.0	0.061	0.962	19.3	0.034	0.997	122.2	0.004
imagenet-a	MetaCLIP-ViT-H/14	confot	0.920	9.6	0.058	0.931	12.7	0.046	0.952	22.8	0.026	0.996	136.8	0.002
sun397	CLIP-ViT-L/14	none	0.923	5.1	0.973	0.931	5.6	0.946	0.953	8.1	0.815	0.997	101.5	0.069
sun397	CLIP-ViT-L/14	confot	0.925	4.4	0.063	0.932	4.9	0.056	0.953	7.1	0.039	0.996	82.7	0.003
sun397	MetaCLIP-ViT-H/14	none	0.926	3.1	0.057	0.935	3.5	0.047	0.955	4.8	0.027	0.997	61.8	0.001
sun397	MetaCLIP-ViT-H/14	confot	0.926	3.0	0.062	0.935	3.4	0.052	0.956	4.7	0.029	0.996	58.4	0.002
aircraft	CLIP-ViT-L/14	none	0.926	10.9	0.904	0.953	15.6	0.682	0.973	20.7	0.487	0.991	32.8	0.173
aircraft	CLIP-ViT-L/14	confot	0.927	10.7	0.107	0.951	14.8	0.080	0.971	20.4	0.058	0.990	33.2	0.027
aircraft	MetaCLIP-ViT-H/14	none	0.924	5.4	0.240	0.951	8.3	0.186	0.972	11.7	0.133	0.991	20.5	0.054
aircraft	MetaCLIP-ViT-H/14	confot	0.925	4.4	0.084	0.949	6.4	0.067	0.968	9.4	0.051	0.987	17.4	0.025
eurosat	CLIP-ViT-L/14	none	0.932	3.8	0.900	0.946	4.4	0.772	0.965	5.1	0.626	0.996	8.4	0.085
eurosat	CLIP-ViT-L/14	confot	0.931	3.4	0.003	0.945	4.0	0.002	0.964	4.6	0.002	0.995	7.4	0.000
eurosat	MetaCLIP-ViT-H/14	none	0.928	6.0	0.001	0.942	6.7	0.001	0.962	7.6	0.000	0.994	9.6	0.000
eurosat	MetaCLIP-ViT-H/14	confot	0.927	3.3	0.000	0.941	3.8	0.000	0.960	4.3	0.000	0.994	7.3	0.000
stanford_cars	CLIP-ViT-L/14	none	0.930	181.9	0.947	0.939	183.7	0.865	0.960	188.0	0.726	0.996	195.2	0.076
stanford_cars	CLIP-ViT-L/14	confot	0.929	181.8	0.000	0.939	183.8	0.000	0.960	187.9	0.000	0.996	195.2	0.000
stanford_cars	MetaCLIP-ViT-H/14	none	0.929	182.4	0.000	0.939	184.4	0.000	0.961	188.6	0.000	0.998	195.6	0.000
stanford_cars	MetaCLIP-ViT-H/14	confot	0.930	182.2	0.000	0.939	184.1	0.000	0.961	188.5	0.000	0.998	195.6	0.000

Table 21. Zero-shot generalization on Food101, Pets, Flowers, Caltech101, DTD, and UCF.

Dataset	Backbone	Adapt	Split CP (APS)			BQ mean			BQ mix95/05			BQ $\mathrm{CVaR}_{0.9}$		
			Cov	Size	$\mathrm{CVaR}_{0.95}$	Cov	Size	CVaR	Cov	Size	CVaR	Cov	Size	CVaR
food101	CLIP-ViT-L/14	none	0.940	1.3	0.079	0.947	1.4	0.070	0.955	1.7	0.060	0.995	19.2	0.006
food101	CLIP-ViT-L/14	confot	0.937	1.4	0.078	0.943	1.6	0.071	0.954	2.1	0.059	0.995	23.4	0.006
food101	MetaCLIP-ViT-H/14	none	0.946	1.2	0.065	0.952	1.3	0.059	0.960	1.6	0.051	0.995	20.1	0.008
food101	MetaCLIP-ViT-H/14	confot	0.942	1.2	0.079	0.949	1.4	0.070	0.957	1.7	0.061	0.995	22.5	0.010
oxford_pets	CLIP-ViT-L/14	none	0.962	1.3	0.007	0.970	1.6	0.005	0.976	1.7	0.004	0.990	3.4	0.001
oxford_pets	CLIP-ViT-L/14	confot	0.959	1.3	0.007	0.967	1.8	0.005	0.973	2.1	0.003	0.989	5.3	0.001
oxford_pets	MetaCLIP-ViT-H/14	none	0.974	1.2	0.002	0.977	1.4	0.002	0.981	1.5	0.001	0.992	2.7	0.000
oxford_pets	MetaCLIP-ViT-H/14	confot	0.973	1.2	0.009	0.977	1.4	0.007	0.980	1.5	0.005	0.991	2.9	0.002
flowers	CLIP-ViT-L/14	none	0.929	6.1	0.081	0.951	11.5	0.043	0.975	21.3	0.022	0.993	37.0	0.007
flowers	CLIP-ViT-L/14	confot	0.928	3.7	0.181	0.950	6.1	0.132	0.975	12.7	0.067	0.994	23.3	0.023
flowers	MetaCLIP-ViT-H/14	none	0.927	2.9	0.078	0.949	7.8	0.048	0.973	17.7	0.012	0.995	35.8	0.002
flowers	MetaCLIP-ViT-H/14	confot	0.932	1.7	0.117	0.952	3.5	0.081	0.974	12.9	0.044	0.995	30.0	0.004
caltech	CLIP-ViT-L/14	none	0.968	1.1	0.034	0.973	1.3	0.029	0.981	2.0	0.020	0.994	3.7	0.006
caltech	CLIP-ViT-L/14	confot	0.971	1.1	0.036	0.976	1.2	0.032	0.982	2.1	0.023	0.995	5.5	0.004
caltech	MetaCLIP-ViT-H/14	none	0.987	1.0	0.014	0.989	1.1	0.013	0.991	1.4	0.009	0.996	2.3	0.003
caltech	MetaCLIP-ViT-H/14	confot	0.986	1.0	0.033	0.987	1.1	0.031	0.990	1.4	0.026	0.995	2.6	0.011
dtd	CLIP-ViT-L/14	none	0.939	14.8	0.005	0.964	21.1	0.001	0.979	25.6	0.000	0.984	28.6	0.000
dtd	CLIP-ViT-L/14	confot	0.937	12.5	0.021	0.965	18.7	0.011	0.978	22.3	0.006	0.984	25.7	0.005
dtd	MetaCLIP-ViT-H/14	none	0.935	5.2	0.009	0.965	10.7	0.002	0.978	15.8	0.001	0.984	19.3	0.000
dtd	MetaCLIP-ViT-H/14	confot	0.937	4.5	0.044	0.967	7.8	0.021	0.980	10.6	0.010	0.986	13.1	0.006
ucf	CLIP-ViT-L/14	none	0.934	4.8	0.031	0.954	8.5	0.021	0.973	13.5	0.012	0.994	28.1	0.001
ucf	CLIP-ViT-L/14	confot	0.932	5.0	0.093	0.952	9.1	0.070	0.973	15.7	0.042	0.994	34.7	0.010
ucf	MetaCLIP-ViT-H/14	none	0.925	2.3	0.064	0.947	5.1	0.046	0.969	9.6	0.025	0.993	24.4	0.004
ucf	MetaCLIP-ViT-H/14	confot	0.930	2.3	0.094	0.952	5.7	0.072	0.972	12.6	0.047	0.993	37.4	0.009

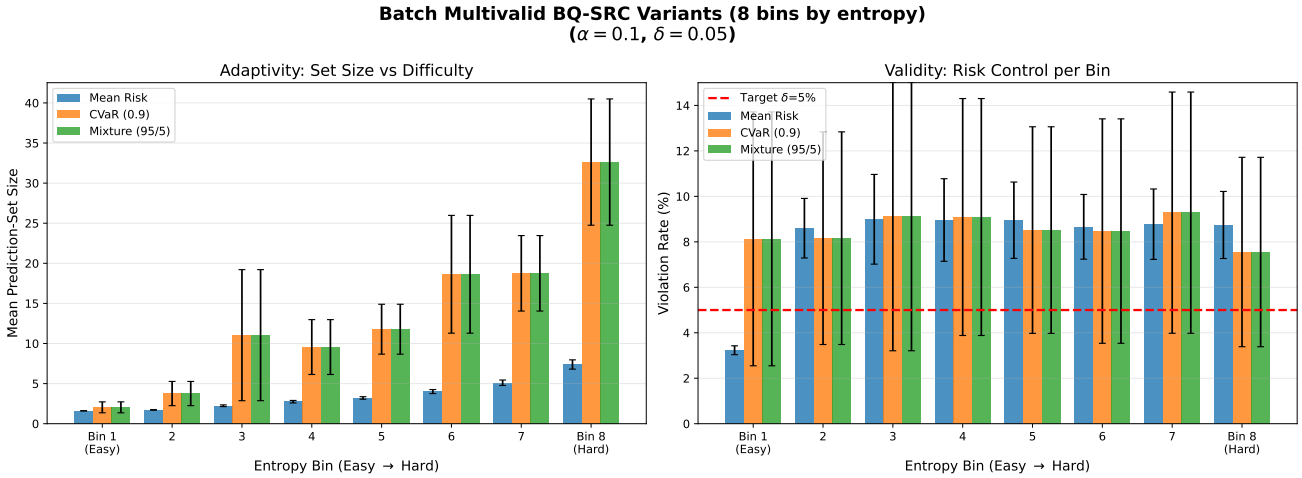

Figure 6. Batch multivalid [17] BQ-SRC variants on MS-COCO ($\alpha = 0.1, \delta = 0.05$). We compare Mean Risk (BQC, blue), $\text{CVaR}_{0.9}$ (orange), and Mixture 95/5 (green) across 8 entropy bins (Easy $\rightarrow$ Hard). **Left:** Mean prediction-set size. All methods adapt by increasing set sizes for harder examples, but CVaR and Mixture variants require significantly larger sets in the hardest bins to satisfy their stricter tail-risk constraints. **Right:** Violation rate. All methods maintain valid risk control (violation rates consistent with $\delta = 5\%$) across all difficulty bins, demonstrating the robustness of the multivalid calibration. The large error bars reflect the smaller sample sizes within each bin.

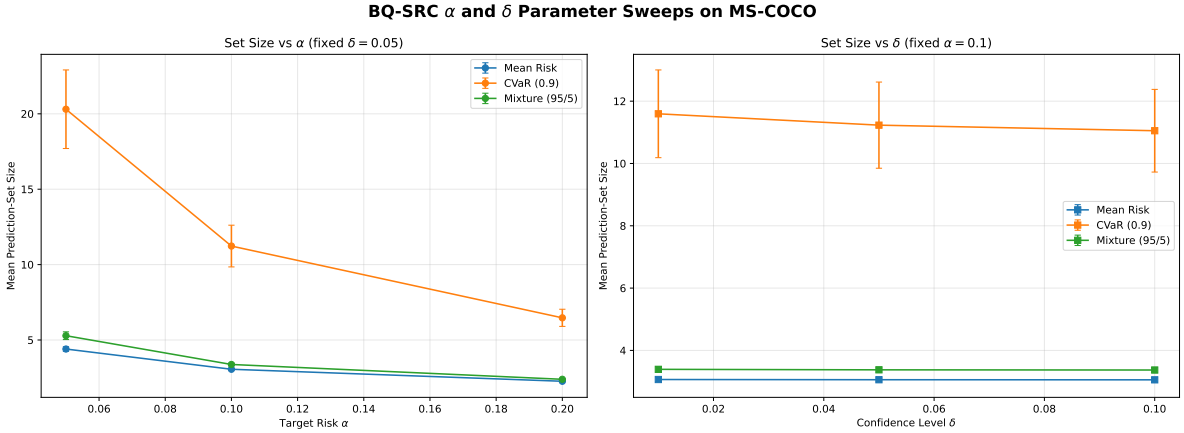

Figure 7. Parameter sweeps for $\alpha \in \{0.05, 0.1, 0.2\}$ and $\delta \in \{0.05, 0.1\}$ on MS-COCO ($M = 5000$, $\gamma = 10^{-3}$). Top: set sizes vs parameters. Bottom: violation rate and set size heatmaps showing monotonic behavior and validity control.

Table 22. ADE20K binary loss with $\alpha = 0.10$ and minimum coverage 0.75. Mean BQ-SRC closely matches CRC (empirical risk ≈ 0.095, set size ≈ 1.5), the mix 95/5 spectrum boosts coverage by ~ 5 pp at the cost of much larger sets, and CVaR remains fully conservative.

Method	Spectral	Multimask	Empirical risk	Coverage	Mean set size
bqsrc	cvar	APS	0.0000 ± 0.0000	0.9998 ± 0.0000	134.062 ± 0.031
bqsrc	mean	APS	0.0940 ± 0.0081	0.9119 ± 0.0054	1.577 ± 0.079
bqsrc	mixture	APS	0.0415 ± 0.0179	0.9587 ± 0.0178	41.948 ± 39.381
crc	mean	APS	0.0985 ± 0.0071	0.9074 ± 0.0048	1.507 ± 0.057
bqsrc	cvar	LAC	0.0000 ± 0.0000	1.0000 ± 0.0000	150.000 ± 0.000
bqsrc	mean	LAC	0.0919 ± 0.0088	0.9128 ± 0.0060	1.510 ± 0.074
bqsrc	mixture	LAC	0.0379 ± 0.0205	0.9604 ± 0.0212	61.141 ± 47.401
crc	mean	LAC	0.0977 ± 0.0072	0.9085 ± 0.0047	1.454 ± 0.047

Table 23. ADE20K miscoverage loss with $\alpha = 0.20$. The relaxed target lets mean-risk methods achieve the goal with ≈ 1 class per pixel; mix spectra offer marginal coverage gains, whereas CVaR stays maximally conservative.

Method	Spectral	Multimask	Empirical risk	Coverage	Mean set size
bqsrc	cvar	APS	0.0002 ± 0.0000	0.9998 ± 0.0000	134.062 ± 0.031
bqsrc	mean	APS	0.1803 ± 0.0010	0.8197 ± 0.0010	1.000 ± 0.000
bqsrc	mixture	APS	0.1777 ± 0.0030	0.8223 ± 0.0030	1.010 ± 0.008
crc	mean	APS	0.1803 ± 0.0010	0.8197 ± 0.0010	1.000 ± 0.000
bqsrc	cvar	LAC	0.0000 ± 0.0000	1.0000 ± 0.0000	150.000 ± 0.000
bqsrc	mean	LAC	0.1803 ± 0.0010	0.8197 ± 0.0010	1.000 ± 0.000
bqsrc	mixture	LAC	0.1776 ± 0.0030	0.8224 ± 0.0030	1.006 ± 0.005
crc	mean	LAC	0.1803 ± 0.0010	0.8197 ± 0.0010	1.000 ± 0.000

Table 24. Cityscapes binary loss — aggregate view across all α and coverage targets. Spectral risk (mean / $\text{CVaR}_{0.9}$ / mix 95/5) is shown for each experiment.

α	τ	Method	Spectral Risk	Empirical Risk	AR
0.10	0.99	CRC	0.082	0.082 (0.020)	1.294 (0.017)
		BQ-SRC (Mean)	0.080	0.080 (0.019)	1.297 (0.014)
		BQ-SRC (CVaR)	0.000	0.000 (0.000)	19.000 (0.000)
		BQ-SRC (Mix)	0.064	0.044 (0.016)	1.456 (0.054)
0.10	0.95	CRC	0.102	0.102 (0.013)	1.035 (0.003)
		BQ-SRC (Mean)	0.093	0.093 (0.015)	1.037 (0.003)
		BQ-SRC (CVaR)	0.016	0.002 (0.003)	6.805 (8.418)
		BQ-SRC (Mix)	0.086	0.059 (0.011)	1.050 (0.005)
0.10	0.90	CRC	0.036	0.036 (0.005)	1.000 (0.000)
		BQ-SRC (Mean)	0.036	0.036 (0.005)	1.000 (0.000)
		BQ-SRC (CVaR)	0.016	0.002 (0.004)	1.402 (0.167)
		BQ-SRC (Mix)	0.052	0.036 (0.005)	1.000 (0.000)
0.10	0.75	CRC	0.003	0.003 (0.003)	1.000 (0.000)
		BQ-SRC (Mean)	0.003	0.003 (0.003)	1.000 (0.000)
		BQ-SRC (CVaR)	0.016	0.002 (0.003)	1.040 (0.043)
		BQ-SRC (Mix)	0.005	0.003 (0.003)	1.000 (0.000)
0.01	0.99	CRC	0.000	0.000 (0.000)	19.000 (0.000)
		BQ-SRC (Mean)	0.000	0.000 (0.000)	19.000 (0.000)
		BQ-SRC (CVaR)	0.000	0.000 (0.000)	19.000 (0.000)
		BQ-SRC (Mix)	0.000	0.000 (0.000)	19.000 (0.000)
0.01	0.95	CRC	0.007	0.007 (0.007)	1.272 (0.162)
		BQ-SRC (Mean)	0.002	0.002 (0.003)	6.805 (8.418)
		BQ-SRC (CVaR)	0.016	0.002 (0.003)	6.805 (8.418)
		BQ-SRC (Mix)	0.002	0.002 (0.003)	6.805 (8.418)
0.01	0.90	CRC	0.007	0.007 (0.005)	1.130 (0.127)
		BQ-SRC (Mean)	0.002	0.002 (0.004)	1.402 (0.167)
		BQ-SRC (CVaR)	0.016	0.002 (0.004)	1.402 (0.167)
		BQ-SRC (Mix)	0.002	0.002 (0.004)	1.402 (0.167)
0.01	0.75	CRC	0.003	0.003 (0.003)	1.007 (0.021)
		BQ-SRC (Mean)	0.002	0.002 (0.003)	1.040 (0.043)
		BQ-SRC (CVaR)	0.016	0.002 (0.003)	1.040 (0.043)
		BQ-SRC (Mix)	0.002	0.002 (0.003)	1.040 (0.043)

Table 25. Cityscapes miscoverage loss — aggregate view across all α values. Spectral risk reflects the specified spectrum per method, recomputed from per-image losses.

α	Method	Spectral Risk	Empirical Risk	AR
0.20	CRC	0.041	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (Mean)	0.041	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (CVaR)	0.106	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (Mix)	0.044	0.041 (0.001)	1.000 (0.000)
0.10	CRC	0.041	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (Mean)	0.041	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (CVaR)	0.089	0.032 (0.007)	1.024 (0.021)
	BQ-SRC (Mix)	0.044	0.041 (0.001)	1.000 (0.000)
0.05	CRC	0.041	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (Mean)	0.041	0.041 (0.001)	1.000 (0.000)
	BQ-SRC (CVaR)	0.040	0.011 (0.004)	1.152 (0.057)
	BQ-SRC (Mix)	0.044	0.041 (0.001)	1.000 (0.000)
0.01	CRC	0.006	0.006 (0.001)	1.232 (0.014)
	BQ-SRC (Mean)	0.009	0.009 (0.001)	1.170 (0.018)
	BQ-SRC (CVaR)	0.002	0.000 (0.001)	15.545 (7.284)
	BQ-SRC (Mix)	0.008	0.007 (0.001)	1.207 (0.032)
0.005	CRC	0.000	0.000 (0.000)	19.000 (0.000)
	BQ-SRC (Mean)	0.004	0.004 (0.001)	1.345 (0.049)
	BQ-SRC (CVaR)	0.000	0.000 (0.000)	19.000 (0.000)
	BQ-SRC (Mix)	0.004	0.003 (0.001)	1.447 (0.072)

Table 26. Cityscapes semantic segmentation results (10 trials)

Setup	Spectral risk	Violations	Prediction-set size
CRC $\alpha = 0.01, \gamma = 0.95$	0.0157 (0.0085, 0.0229)	0.00% (0.00, 30.85)	1.2549 (1.1617, 1.3481)
CRC $\alpha = 0.01, \gamma = 0.99$	0.0162 (0.0039, 0.0286)	0.00% (0.00, 30.85)	2.2912 (2.0898, 2.4926)
CRC $\alpha = 0.10, \gamma = 0.95$	0.1416 (0.1265, 0.1566)	100.00% (69.15, 100.00)	1.0351 (1.0330, 1.0373)
CRC $\alpha = 0.10, \gamma = 0.99$	0.1481 (0.1341, 0.1622)	90.00% (55.50, 99.75)	1.2708 (1.2638, 1.2779)

Table 27. LoveDA semantic segmentation results (10 trials)

Setup	Spectral risk	Violations	Prediction-set size
CRC $\alpha = 0.01, \gamma = 0.50$	0.0144 (0.0060, 0.0227)	0.00% (0.00, 30.85)	3.6680 (3.4363, 3.8996)
CRC $\alpha = 0.01, \gamma = 0.75$	0.0125 (0.0065, 0.0185)	0.00% (0.00, 30.85)	4.9822 (4.7953, 5.1691)
CRC $\alpha = 0.10, \gamma = 0.50$	0.1400 (0.1278, 0.1522)	100.00% (69.15, 100.00)	1.1503 (1.0906, 1.2100)
CRC $\alpha = 0.10, \gamma = 0.75$	0.1471 (0.1367, 0.1574)	100.00% (69.15, 100.00)	2.5448 (2.4162, 2.6734)

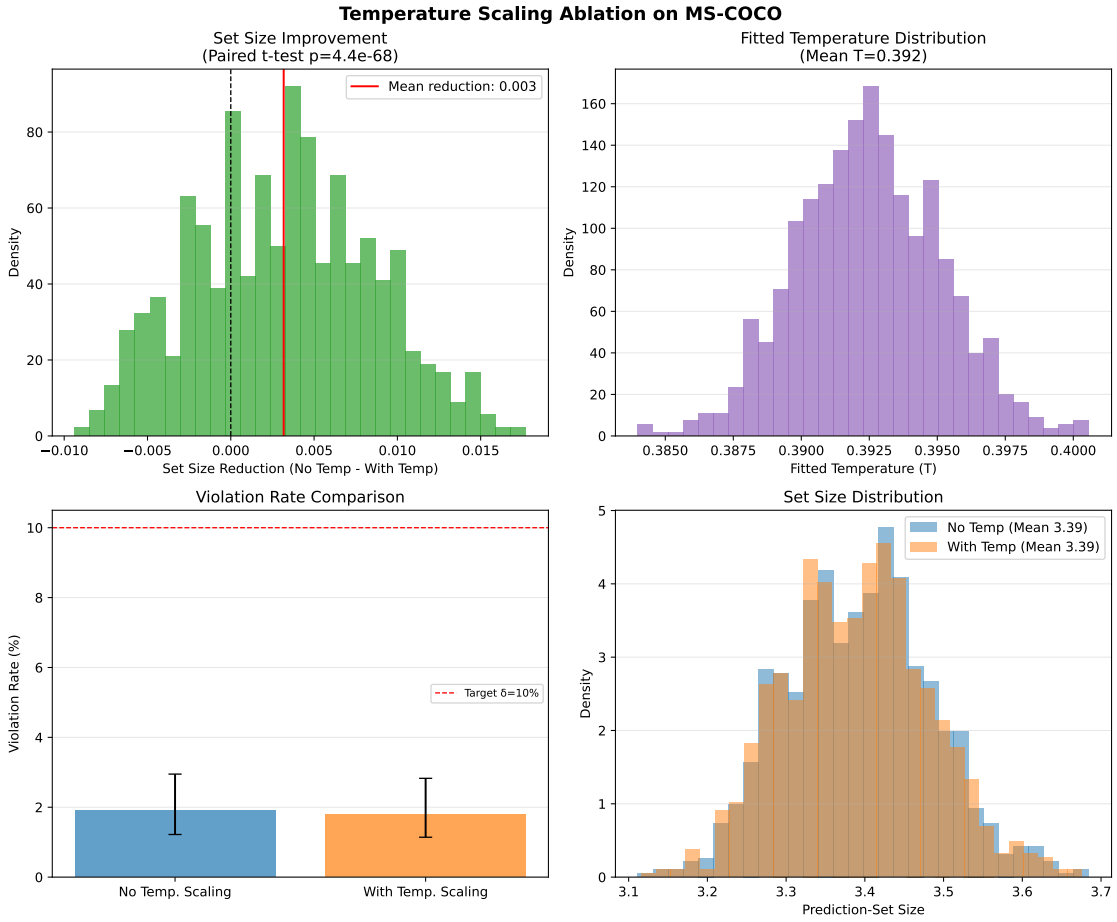

Figure 8. Temperature scaling ablation on MS-COCO. Comparison of BQ-SRC with and without temperature scaling [15] shows minimal reduced set sizes while maintaining validity.