How to Trust Your Diffusion Model: A Convex Optimization Approach to Conformal Risk Control

Jacopo Teneggi [1 2] **Matthew Tivnan** [3] **J. Webster Stayman** [3] **Jeremias Sulam** [3 2]

Abstract

Score-based generative modeling, informally referred to as *diffusion models*, continue to grow in popularity across several important domains and tasks. While they provide high-quality and diverse samples from empirical distributions, important questions remain on the reliability and trustworthiness of these sampling procedures for their responsible use in critical scenarios. Conformal prediction is a modern tool to construct finite-sample, distribution-free uncertainty guarantees for any black-box predictor. In this work, we focus on image-to-image regression tasks and we present a generalization of the Risk-Controlling Prediction Sets (RCPS) procedure, that we term K-RCPS, which allows to *(i)* provide entrywise calibrated intervals for future samples of any diffusion model, and *(ii)* control a certain notion of risk with respect to a ground truth image with minimal mean interval length. Differently from existing conformal risk control procedures, ours relies on a novel convex optimization approach that allows for multidimensional risk control while provably minimizing the mean interval length. We illustrate our approach on two real-world image denoising problems: on natural images of faces as well as on computed tomography (CT) scans of the abdomen, demonstrating state of the art performance.

1. Introduction

Generative modeling is one of the longest standing tasks of classical and modern machine learning (Bishop & Nasrabadi, 2006). Recently, the foundational works by Song & Ermon (2019); Song et al. (2020a); Pang et al. (2020) on sampling via *score-matching* (Hyvärinen & Dayan, 2005) and by Ho et al. (2020) on *denoising diffusion models* (Sohl-Dickstein et al., 2015) paved the way for a new class of *score-based* generative models, which solve a reverse-time stochastic differential equation (SDE) (Song et al., 2020b; Anderson, 1982). These models have proven remarkably effective on both unconditional (i.e., starting from random noise) and conditional (e.g., inpainting, denoising, super-resolution, or class-conditional) sample generation across a variety of fields (Yang et al., 2022; Croitoru et al., 2022). For example, score-based generative models have been applied to inverse problems in general computer vision and medical imaging (Kadkhodaie & Simoncelli, 2021; Kawar et al., 2021b;a; Xie & Li, 2022; Torem et al., 2022; Song et al., 2021), 3D shape generation (Zeng et al., 2022; Xu et al., 2022; Metzer et al., 2022), and even in protein design (Hoogeboom et al., 2022; Corso et al., 2022; Watson et al., 2022; Ingraham et al., 2022).

These strong empirical results highlight the potential of score-based generative models. However, they currently lack of precise statistical guarantees on the distribution of the generated samples, which hinders their safe deployment in high-stakes scenarios (Horwitz & Hoshen, 2022). For example, consider a radiologist who is shown a computed tomography (CT) scan of the abdomen of a patient reconstructed via a score-based generative model. How confident should they be of the fine-grained details of the presented image? Should they trust that the model has not *hallucinated* some of the features (e.g., calcifications, blood vessels, or nodules) involved in the diagnostic process? Put differently, how different will future samples be from the presented image, and how far can we expect them to be from the ground truth image?

In this work we focus on image-to-image regression problems, where we are interested in recovering a high-quality ground truth image given a low-quality observation. While our approach is general, we focus on the problem of image denoising as a running example. We address the questions posed above on the reliability of score-based generative models (and, more generally, of any sampling procedure)

[1]Department of Computer Science, Johns Hopkins University, Baltimore, USA [2]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, USA [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA. Correspondence to: Jacopo Teneggi <jtenegg1@jhu.edu>, Jeremias Sulam <jsulam1@jhu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

through the lens of conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; Lei & Wasserman, 2014; Shafer & Vovk, 2008; Angelopoulos & Bates, 2021) and conformal risk control (Bates et al., 2021; Angelopoulos et al., 2021; 2022a) which provide any black-box predictor with *distribution-free*, *finite-sample* uncertainty guarantees. In particular, the contribution of this paper is three-fold:

1. Given a fixed score network, a low-quality observation, and any sampling procedure, we show how to construct valid entrywise calibrated intervals that provide *coverage* of future samples, i.e. future samples (on the same observation) will fall within the intervals with high probability;
2. We introduce a novel high-dimensional conformal risk control procedure that minimizes the mean interval length directly, while guaranteeing the number of pixels in the ground truth image that fall outside of these intervals is below a user-specified level on future, unseen low-quality observations;
3. We showcase our approach for denoising of natural face images as well as for computed tomography of the abdomen, achieving state of the art results in mean interval length.

Going back to our example, providing such uncertainty intervals with provable statistical guarantees would improve the radiologist's trust in the sense that these intervals precisely characterize the type of tissue that could be reconstructed by the model. Lastly, even though our contributions are presented in the context of score-based generative modeling for regression problems—given their recent popularity (Kazerouni et al., 2022; Yang et al., 2022; Croitoru et al., 2023)—our results are broadly applicable to any sampling procedure, and we will comment on potential direct extensions where appropriate.

1.1. Related work

Image-to-Image Risk Control Previous works have explored conformal risk control procedures for image-to-image regression tasks. In particular, Angelopoulos et al. (2022b) show how to construct set predictors from heuristic notions of uncertainty (e.g., quantile regression (Koenker & Bassett Jr, 1978; Romano et al., 2019)) for any image regressor, and how to calibrate the resulting intervals according to the original RCPS procedure of Bates et al. (2021). Kutiel et al. (2022) move beyond set predictors and propose a mask-based conformal risk control procedure that allows for notions of distance between the ground truth and predicted images other than interval-based ones. Finally, and most closely to this paper, Horwitz & Hoshen (2022) sketch ideas of conformal risk control for diffusion models with the intention to integrate quantile regression and produce heuristic sampling bounds without the need to sample several times. Horwitz & Hoshen (2022) also use the original RCPS procedure to guarantee risk control. Although similar in spirit, the contribution of this paper focuses on a high-dimensional generalization of the original RCPS procedure that formally minimizes the mean interval length. Our proposed procedure is agnostic of the notion of uncertainty chosen to construct the necessary set predictors.

2. Background

First, we briefly introduce the necessary notation and general background information. Herein, we will refer to images as vectors in $\mathbb{R}^d$, such that $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^d$ indicate the space of high-quality ground truth images, and low-quality observations, respectively. We assume both $\mathcal{X}$ and $\mathcal{Y}$ to be bounded. For a general image-to-image regression problem, given a pair (x, y) drawn from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the task is to retrieve $x \in \mathcal{X}$ given $y \in \mathcal{Y}$. This is usually carried out by means of a predictor $f : \mathcal{Y} \to \mathcal{X}$ that minimizes some notion of distance (e.g., MSE loss) between the ground truth images and reconstructed estimates on a set $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ of n pairs of high- and low-quality images. For example, in the classical denoising problem, one has $y = x + v_0$ where $v_0 \sim \mathcal{N}(0, \sigma_0^2 \mathbb{I})$ is random Gaussian noise with variance σ_0^2, and one wishes to learn a denoiser f such that $f(y) \approx x$.

2.1. Score-based Conditional Sampling

Most image-to-image regression problems are ill-posed: there exist several ground truth images that could have generated the same low-quality observation. This is easy to see for the classical denoising problem described above. Instead of a point predictor f—which could approximate a maximum-a-posteriori (MAP) estimate—one is often interested in devising a sampling procedure $F : \mathcal{Y} \to \mathcal{X}$ for the posterior $p(x|y)$, which precisely describes the distribution of possible ground truth images that generated the observation y. In real-world scenarios, however, the full joint (x, y) is unknown, and one must resort to approximate $p(x|y)$ from finite data. It is known that for a general Itô process $\mathrm{d}x = h(x, t)\,\mathrm{d}t + g(t)\,\mathrm{d}w$ that perturbs an input x into random noise (Karatzas et al., 1991), it suffices to know the *Stein score* $\nabla_x \log p_t(x)$ (Anderson, 1982; Liu et al., 2016) to sample from $p(x)$ via the reverse-time process

$$\mathrm{d}x = [h(x, t) - g(t)^2 \nabla_x \log p_t(x)]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{w}, \quad (1)$$

where $h(x, t)$ and $g(t)$ are a *drift* and *diffusion* term, respectively, and $\mathrm{d}w$ and $\mathrm{d}\bar{w}$ are forward- and reverse-time standard Brownian motion.[1] Furthermore, if the likelihood

[1] We will assume time t continuous in $[0, 1]$.

$p(y|x)$ is known—which is usually the case for image-to-image regression problems—it is possible to condition the sampling procedure on an observation y. Specifically, by Bayes' rule, it follows that $\nabla_x \log p_t(x|y) = \nabla_x \log p_t(y|x) + \nabla_x \log p_t(x)$ which can be plugged-in into the reverse-time SDE in Equation (1) to sample from $p(x|y)$.

Recent advances in generative modeling by Song & Ermon (2019); Song et al. (2020b) showed that one can efficiently train a *time-conditional* score network $s(\tilde{x}, t)$ to approximate the score $\nabla_x \log p_t(\tilde{x})$ via *denoising score-matching* (Hyvärinen & Dayan, 2005). In this way, given a forward-time SDE that models the observation process, a score network $s(\tilde{x}, t) \approx \nabla_x \log p_t(\tilde{x})$, and the likelihood term $p(y|\tilde{x})$, one can sample from $p(x|y)$ by solving the conditional reverse-time SDE with any discretization (e.g., Euler-Maruyama) or predictor-corrector scheme (Song et al., 2020b). While these models perform remarkably well in practice, limited guarantees exist on the distributions that they sample from (Lee et al., 2022). Instead, we will provide guarantees for diffusion models by leveraging ideas of conformal prediction and conformal risk control, which we now introduce.

2.2. Conformal Prediction

Conformal prediction has a rich history in mathematical statistics (Vovk et al., 2005; Papadopoulos et al., 2002; Vovk, 2015; Barber et al., 2021; 2022; Gupta et al., 2022).[2] It comprises various methodologies to construct finite-sample, statistically valid uncertainty guarantees for general predictors without making any assumption on the distribution of the response (i.e., they are distribution-free). It particular, these methods construct valid prediction sets that provide coverage, which we now define.

Definition 2.1 (Coverage (Shafer & Vovk, 2008))**.** Let $z_1, \ldots, z_m, z_{m+1}$ be $m+1$ exchangeable random variables drawn from the same unknown distribution $\mathcal{Q}$ over $\mathcal{Z}$. For a desired miscoverage level $\alpha \in [0, 1]$, a set $\mathcal{C} \subseteq 2^{\mathcal{Z}}$ that only depends on $z_1, \ldots, z_m$ provides coverage if

$$\mathbb{P}[z_{m+1} \in \mathcal{C}] \geq 1 - \alpha. \tag{2}$$

We remark that the notion of coverage defined above was introduced in the context of classification problems, where one is interested in guaranteeing that the true, unseen label of a future sample will be in the prediction set $\mathcal{C}$ with high probability. It is immediate to see how conformal prediction conveys a very precise notion of uncertainty—the larger $\mathcal{C}$ has to be in order to guarantee coverage, the more *uncertain* the underlying predictor. We refer the interested reader to (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021) for classical examples of conformal prediction.

[2]Throughout this work, we will refer to *split* conformal prediction (Vovk et al., 2005) simply as *conformal prediction*.

In many scenarios (e.g., regression), the natural notion of uncertainty may be different from miscoverage as described above (e.g., ℓ_2 norm). We now move onto presenting conformal risk control, which extends the coverage to any notion of risk.

2.3. Conformal Risk Control

Let $\mathcal{I} : \mathcal{Y} \to \mathcal{X}'$ be a general *set-valued* predictor from $\mathcal{Y}$ into $\mathcal{X}' \subseteq 2^{\mathcal{X}}$. Consider a nonnegative loss $\ell : \mathcal{X} \times \mathcal{X}' \to \mathbb{R}$ measuring the discrepancy between a ground truth x and the predicted intervals $\mathcal{I}(y)$. We might be interested in guaranteeing that this loss will be below a certain tolerance $\epsilon \geq 0$ with high probability on future, unseen samples y for which we do not know the ground truth x. Conformal risk control (Bates et al., 2021; Angelopoulos et al., 2021; 2022a) extends ideas of conformal prediction in order to select a specific predictor $\mathcal{I}$ that controls the risk $\mathbb{E}[\ell(x, \mathcal{I}(y))]$ in the following sense.

Definition 2.2 (Risk Controlling Prediction Sets)**.** Let $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ be a calibration set of n i.i.d. samples from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. For a desired risk level $\epsilon \geq 0$ and a failure probability $\delta \in [0, 1]$, a random set-valued predictor $\mathcal{I} : \mathcal{Y} \to \mathcal{X}' \subseteq 2^{\mathcal{X}}$ is an (ϵ, δ)-RCPS w.r.t. a loss function $\ell : \mathcal{X} \times \mathcal{X}' \to \mathbb{R}$ if

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}}[\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x, \mathcal{I}(y))] \leq \epsilon] \geq 1 - \delta. \tag{3}$$

Bates et al. (2021) introduced the first conformal risk control procedure for *monotonically nonincreasing* loss functions, those that satisfy, for a fixed x,

$$\mathcal{I}(y) \subset \mathcal{I}'(y) \implies \ell(x, \mathcal{I}'(y)) \leq \ell(x, \mathcal{I}(y)). \tag{4}$$

In this way, increasing the size of the sets cannot increase the value of the loss. Furthermore, assume that for a fixed input y the family of set predictors $\{\mathcal{I}_\lambda(y)\}_{\lambda \in \Lambda}$, indexed by $\lambda \in \Lambda$, $\Lambda \subset \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, satisfies the following nesting property (Gupta et al., 2022)

$$\lambda_1 < \lambda_2 \implies \mathcal{I}_{\lambda_1}(y) \subset \mathcal{I}_{\lambda_2}(y). \tag{5}$$

Denote $R(\lambda) = \mathbb{E}[\ell(x, \mathcal{I}_\lambda(y))]$ the risk of $\mathcal{I}_\lambda(y)$ and $\hat{R}(\lambda)$ its empirical estimate over a calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$. Finally, let $\hat{R}^+(\lambda)$ be a *pointwise upper confidence bound* (UCB) that covers the risk, that is

$$\mathbb{P}[R(\lambda) \leq \hat{R}^+(\lambda)] \geq 1 - \delta \tag{6}$$

for *each*, fixed value of λ—such that can be derived by means of concentration inequalities (e.g., Hoeffding's inequality (Hoeffding, 1994), Bentkus' inequality (Bentkus, 2004), or respective hybridization (Bates et al., 2021)).[3] With these elements, Bates et al. (2021) show that choosing

$$\hat{\lambda} = \inf\{\lambda \in \Lambda : \hat{R}^+(\lambda') < \epsilon, \forall \lambda' \geq \lambda\} \tag{7}$$

[3]We stress that Equation (6) does *not* imply *uniform coverage* $\forall \lambda \in \Lambda$.

guarantees that $\mathcal{I}_{\hat{\lambda}}(y)$ is an (ϵ, δ)-RCPS according to Definition 2.2. In other words, choosing $\hat{\lambda}$ as the smallest λ such that the UCB is below the desired level ϵ for all values of $\lambda \geq \hat{\lambda}$ controls the risk al level ϵ with probability at least $1 - \delta$. For the sake of completeness, we include the original conformal risk control procedure in Algorithm 2 in Appendix B.

Equipped with these general concepts, we now move onto presenting the contributions of this work.

3. How to Trust Your Diffusion Model

We now go back to the main focus of this paper: solving image-to-image regression problems with diffusion models. Rather than a point-predictor $f : \mathcal{Y} \to \mathcal{X}$, we assume to have access to a stochastic sampling procedure $F : \mathcal{Y} \to \mathcal{X}$ such that $F(y)$ is a random variable with unknown distribution $\mathcal{Q}_y$—that hopefully approximates the posterior distribution of x given y, i.e. $\mathcal{Q}_y \approx p(x|y)$. However, we make no assumptions on the quality of this approximation for our results to hold. As described in Section 2.1, F can be obtained by means of a time-conditional score network $s(\tilde{x}, t)$ and a reverse-time SDE. While our results are applicable to *any* sampling procedure, we present them in the context of diffusion models because of their remarkable empirical results and increasing use in critical applications (Yang et al., 2022; Croitoru et al., 2022).

One can identify three separate sources of randomness in a general stochastic image-to-image regression problem: *(i)* the unknown prior $p(x)$ over the space of ground-truth images, as $x \sim p(x)$, *(ii)* the randomness in the observation process of y (which can be modeled by a forward-time SDE over x), and finally *(iii)* the stochasticity in the sampling procedure $F(y)$. We will first provide conformal prediction guarantees for a fixed observation y, and then move onto conformal risk control for the ground truth image x.

3.1. Calibrated Quantiles for Future Samples

Given the same low-quality (e.g., noisy) observation y, where will future unseen samples from $F(y) \sim \mathcal{Q}_y$ fall? How concentrated will they be? Denote $\mathcal{I} : \mathcal{Y} \to \mathcal{X}'$ a (random) set-valued predictor from $\mathcal{Y} \subset \mathbb{R}^d$ into a space of sets $\mathcal{X}' \subseteq 2^{\mathcal{X}}$ over $\mathcal{X} \subset \mathbb{R}^d$ (e.g., $\mathcal{X} = [0,1]^d$, $\mathcal{X}' \subseteq 2^{[0,1]^d}$). We extend the notion of coverage in Definition 2.1 to *entrywise coverage*, which we now make precise.

Definition 3.1 (Entrywise coverage)**.** Let $z_1, \ldots, z_m, z_{m+1}$ be $m+1$ exchangeable random vectors drawn from the same unknown distribution $\mathcal{Q}$ over $\mathcal{X} \subset \mathbb{R}^d$. For a desired miscoverage level $\alpha \in [0,1]$, a set $\mathcal{I} \subseteq 2^{\mathcal{X}}$ that only depends on $z_1, \ldots, z_m$ provides entrywise coverage if

$$\mathbb{P}[(z_{m+1})_j \in \mathcal{I}_j] \geq 1 - \alpha \tag{8}$$

for each $j \in [d] \coloneqq \{1, \ldots, d\}$.

We stress that the definition above is different from notions of *vector quantiles* (Carlier et al., 2016; Chernozhukov et al., 2017) in the sense that coverage is not guaranteed over the entire new random vector z_{m+1} but rather along each dimension independently. Ideas of vector quantile regression (VQR) are complementary to the contribution of the current work and subject of ongoing research (Genevay et al., 2016; Carlier et al., 2020; Rosenberg et al., 2022).

For a fixed observation y, we use conformal prediction to construct a set predictor that provides entrywise coverage.

Lemma 3.2 (Calibrated quantiles guarantee entrywise coverage)**.** *Let $F : \mathcal{Y} \to \mathcal{X}$ be a stochastic sampling procedure from $\mathcal{Y} \subset \mathbb{R}^d$ into $\mathcal{X} \subset \mathbb{R}^d$. Given $y \in \mathcal{Y}$, let $F_1, \ldots, F_m, F_{m+1}$ be $m+1$ i.i.d. samples from $F(y)$. For a desired miscoverage level $\alpha \in [0,1]$ and for each $j \in [d]$, let $\hat{l}_{j,\alpha}, \hat{u}_{j,\alpha}$ be the $\lfloor (m+1)\alpha/2 \rfloor / m$ and $\lceil (m+1)(1-\alpha/2) \rceil / m$ entrywise calibrated empirical quantiles of $F_1, \ldots, F_m$. Then,*

$$\mathcal{I}^{\alpha}(y)_j = [\hat{l}_{j,\alpha}, \hat{u}_{j,\alpha}] \tag{9}$$

provides entrywise coverage.

The simple proof of this result is included in Appendix A.1. We remark that, analogously to previous works (Angelopoulos et al., 2022b; Horwitz & Hoshen, 2022), the intervals in $\mathcal{I}^{\alpha}(y)$ are *feature-dependent* and they capture regions of the image where the sampling process $F(y)$ may have larger uncertainty. The intervals in $\mathcal{I}^{\alpha}(y)$ are statistically valid for any number of samples m and any distribution $\mathcal{Q}_y$, i.e. they are not a heuristic notion of uncertainty. If the sampling procedure F is a diffusion model, constructing $\mathcal{I}^{\alpha}(y)$ is agnostic of the discretization scheme used to solve the reverse-time SDE (Song et al., 2020b) and it does not require retraining the underlying score network, which can be a time-consuming and delicate process, especially when the size of the images is considerable. On the other hand, constructing the intervals $\mathcal{I}^{\alpha}(y)$ requires sampling a large enough number of times from $F(y)$, which may seen cumbersome (Horwitz & Hoshen, 2022). This is by construction and intention: diffusion models are indeed very useful in providing good (and varied) samples from the approximate posterior. In this way, practitioners do typically sample several realizations to get an empirical study of this distribution. In these settings, constructing the intervals $\mathcal{I}^{\alpha}(y)$ does not involve any additional computational costs. Furthermore, note that sampling is completely parallelizable, and so no extra complexity is incurred if a larger number of computing nodes are available.

3.2. A Provable Approach to Optimal Risk Control

In this section, we will revisit the main ideas around conformal risk control introduced in Section 2.3 and generalize them into our proposed approach, K-RCPS. Naturally, one would like a good conformal risk control procedure to yield the shortest possible interval lengths. Assume pixel intensities are normalized between $[0, 1]$ and consider the loss function

$$\ell^{01}(x, \mathcal{I}(y)) = \frac{1}{d} \sum_{j \in [d]} \mathbb{1}[x_j \notin \mathcal{I}(y)_j], \tag{10}$$

which counts the (average) number of ground truth pixels that fall outside of their respective intervals in $\mathcal{I}(y)$. The constant set-valued predictor $\mathcal{U}(y) = [0, 1]^d$ would trivially control the risk, i.e. $R^{01}(\lambda) = \mathbb{E}[\ell^{01}(x, \mathcal{U}(y))] = 0$. Alas, such a predictor would be completely uninformative. Instead, let $\{\mathcal{I}_\lambda(y)\}_{\lambda \in \Lambda}$, $\Lambda \subset \overline{\mathbb{R}}$ be a family of predictors that satisfies the nesting property in Equation (5). In particular, we propose the following additive parametrization in λ

$$\mathcal{I}_\lambda(y)_j = [\hat{l}_j - \lambda, \hat{u}_j + \lambda] \tag{11}$$

for some lower and upper endpoints $\hat{l}_j < \hat{u}_j$ that may depend on y. For this particularly chosen family of nested predictors, it follows that the mean interval length is

$$\bar{I}(\lambda) = \frac{1}{d} \sum_{j \in [d]} (\hat{u}_j - \hat{l}_j) + 2\lambda, \tag{12}$$

a linear function of λ. Moreover, we can instantiate $\hat{l}_j$ and $\hat{u}_j$ to be the calibrated quantiles with entrywise coverage, i.e. $\mathcal{I}^\alpha_\lambda(y) = [\hat{l}_{j,\alpha} - \lambda, \hat{u}_{j,\alpha} + \lambda]$.

For such a class of predictors—since the ℓ^{01} loss is monotonically nonincreasing—the original RCPS procedure (see Equation (7)) is equivalent to the following constrained optimization problem

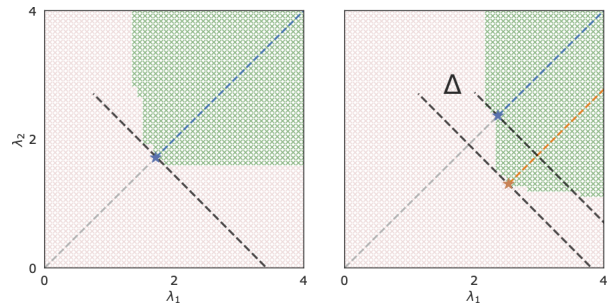
$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \bar{I}(\lambda) \quad \text{s.t.} \quad \hat{R}^{01+}(\lambda') < \epsilon, \ \forall \lambda' \geq \lambda \quad (\mathrm{P}_1)$$

which naturally minimizes λ. However, optimizing the mean interval length over a single scalar parameter λ is suboptimal in general, as shown in Figure 1. With abuse of notation—we do not generally refer to vectors with boldface—let $\{\mathcal{I}_{\boldsymbol{\lambda}}(y)\}_{\boldsymbol{\lambda} \in \Lambda^d}$ be a family of predictors indexed by a d-dimensional *vector* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ that satisfies the nesting property in Equation (5) in an entrywise fashion. A natural extension of Equation (11) is then

$$\mathcal{I}_{\boldsymbol{\lambda}}(y)_j = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j], \tag{13}$$

from which one can define an equivalent function $\bar{I}(\boldsymbol{\lambda})$. In particular, using the calibrated intervals as before, define

$$\mathcal{I}^\alpha_{\boldsymbol{\lambda}}(y) = [\hat{l}_{j,\alpha} - \lambda_j, \hat{u}_{j,\alpha} + \lambda_j]. \tag{14}$$



(a) $\mu = [-1, 1]^T$. (b) $\mu = [-2, 0.75]^T$.

Figure 1: Pictorial representation of the suboptimality of the choice of a single scalar parameter λ w.r.t. the mean interval length. $\mathcal{S}_{\text{cal}} \sim \mathcal{N}(\mu, \mathbb{I}_2)^n$, $n = 128$, and $(\mathcal{I}_{\boldsymbol{\lambda}})_j = [-1 - \lambda_j, 1 + \lambda_j]$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. For $\epsilon = \delta = 0.1$, $\hat{R}^{01+}(\boldsymbol{\lambda})$ is obtained via Hoeffding-Bentkus hybridization. Green areas indicate regions where $\hat{R}^{01+}(\boldsymbol{\lambda}) \leq \epsilon$, and conversely for red regions. (a) Shows that when features are concentrated symmetrically around the intervals, minimizing $\lambda_1 = \lambda_2 = \lambda$ (blue star) minimizes the mean interval length, while (b) shows that in the general case, the optimal $\boldsymbol{\lambda}$ (orange star) may have $\lambda_1 \neq \lambda_2$. Δ highlights the gain in mean interval length obtained by choosing the orange star instead of the blue one.

Note now that $\ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ is entrywise monotonically nonincreasing. Hence, for a fixed vector $\boldsymbol{\eta} \in \mathbb{R}^d$ in the positive orthant (i.e., $\boldsymbol{\eta} \geq 0$, entrywise), the d-dimensional extension of (P_1) becomes

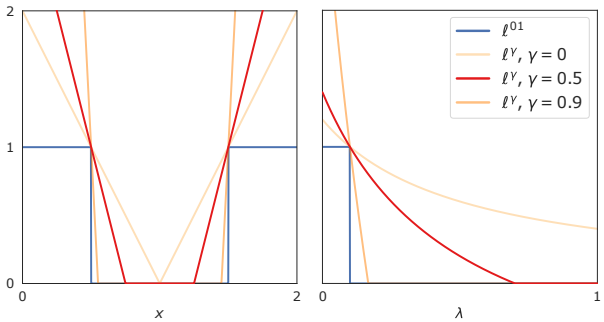
$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda^d} \sum_{j \in [d]} \lambda_j \quad \text{s.t.} \quad \hat{R}^{01+}(\boldsymbol{\lambda} + \beta \boldsymbol{\eta}) < \epsilon, \quad (\mathrm{P}_d)$$

$\forall \beta \geq 0$. We include an explicit analytical expression for $\hat{R}^{01+}(\boldsymbol{\lambda} + \boldsymbol{\eta})$ in Appendix C. Intuitively, $\hat{\boldsymbol{\lambda}}$ minimizes the sum of its entries such that the UCB is smaller than ϵ for all points *to its right* along the direction of $\boldsymbol{\eta}$ parametrized by β. We now show a general high-dimensional risk control result that holds for any entrywise monotonically nonincreasing loss function ℓ (and not just ℓ^{01} as presented in (P_d)) with risk $R(\boldsymbol{\lambda})$, empirical estimate $\hat{R}(\boldsymbol{\lambda})$ and respective UCB $\hat{R}^+(\boldsymbol{\lambda})$.

Theorem 3.3 (Optimal mean interval length risk control)**.** *Let $\ell : \mathcal{X} \times \mathcal{X}' \to \mathbb{R}$, $\mathcal{X}' \subseteq 2^{\mathcal{X}}$, $\mathcal{X} \subset \mathbb{R}^d$ be an entrywise monotonically nonincreasing function and let $\{\mathcal{I}_{\boldsymbol{\lambda}}(y) = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j]\}_{\boldsymbol{\lambda} \in \Lambda^d}$ be a family of set-valued predictors $\mathcal{I} : \mathcal{Y} \to \mathcal{X}'$, $\mathcal{Y} \subset \mathbb{R}^d$ indexed by $\boldsymbol{\lambda} \in \Lambda^d$, $\Lambda \subset \overline{\mathbb{R}}$, for some lower and upper bounds $\hat{l}_j < \hat{u}_j$ that may depend on y. For a fixed vector $\boldsymbol{\eta} \in \mathbb{R}^d$, $\boldsymbol{\eta} \geq 0$, if*

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda^d} \sum_{j \in [d]} \lambda_j \quad \textit{s.t.} \quad \hat{R}^+(\boldsymbol{\lambda} + \beta \boldsymbol{\eta}) < \epsilon \tag{15}$$

$\forall \beta \geq 0$, then $\mathcal{I}_{\hat{\boldsymbol{\lambda}}}(y)$ is an (ϵ, δ)-RCPS, and $\hat{\boldsymbol{\lambda}}$ minimizes the mean interval length.



(a) ℓ^{01}, ℓ^{γ} as a function of x. (b) ℓ^{01}, ℓ^{γ} as a function of λ.

Figure 2: Visualization of $\ell^{01}(x, \mathcal{I}_\lambda(y))$ and $\ell^{\gamma}(x, \mathcal{I}_\lambda(y))$ for $\mathcal{I}_\lambda(y) = [0.50 - \lambda, 1.50 + \lambda]$, $\gamma \in \{0, 0.5, 0.9\}$. In (a) $\lambda = 0$, and in (b) $x = 1.6$.

The proof is included in Appendix A.2. Since ℓ^{01} is entrywise monotonically nonincreasing, it follows that the solution to (P_d) controls risk. The attentive reader will have noticed (as shown in Figure 1) that the constraint set $\hat{R}^{01+}(\boldsymbol{\lambda}) \leq \epsilon$ need not be convex. Furthermore, and as shown in Figure 2b, ℓ^{01} is not convex in $\boldsymbol{\lambda}$. Hence, it is not possible to optimally solve (P_d) directly. Instead, we relax it to a convex optimization problem by means of a convex upper bound[4]

$$\ell^{\gamma}(x, \mathcal{I}_{\boldsymbol{\lambda}}(y)) = \frac{1}{d} \sum_{j \in [d]} \left[\frac{2(1+q)}{I(\boldsymbol{\lambda})_j} |x_j - c_j| - q \right]_+, \quad (16)$$

where $q = \gamma/(1-\gamma)$, $\gamma \in [0, 1)$, $I(\boldsymbol{\lambda})_j = \hat{u}_j - \hat{l}_j + 2\lambda_j$, $c_j = (\hat{u}_j + \hat{l}_j)/2$, and $[\cdot]_+ = \max(0, \cdot)$. As shown in Figure 2a, the hyperparameter γ controls the degree of relaxation by means of changing the portion of the intervals $[\hat{l}_j, \hat{u}_j]$ where the loss is 0. This way, $\gamma = 0$ retrieves the ℓ_1 loss centered at c_j, and $\lim_{\gamma \to 1} \ell^{\gamma} = \infty$ if $\exists j \in [d] : \; x_j \notin [\hat{l}_j, \hat{u}_j]$ and 0 otherwise.

While one can readily propose a convex alternative to (P_d) by means of this new loss, we instead propose a generalization of this idea in our final problem formulation

$$\widetilde{\boldsymbol{\lambda}}_K = \underset{\boldsymbol{\lambda} \in \Lambda^K}{\arg\min} \sum_{k \in [K]} n_k \lambda_k \quad \text{s.t.} \quad \hat{R}^{\gamma}(M\boldsymbol{\lambda}) \leq \epsilon, \quad (\mathrm{P}_K)$$

for any user-defined K-partition of the $[d]$ features—which can be identified by a membership matrix $M \in \{0, 1\}^{d \times K}$ where each feature belongs to (only) one of the K groups with $n_k := |\{j \in [d] : \; M_{jk} = 1\}|$, $\sum_{k \in [K]} n_k = d$. As we will shortly see, it will be useful to define these groups as the empirical quantiles; i.e. set M as that assigning each feature to their respective k^{th} quantile of the entrywise empirical loss over the optimization set (which is a vector in $\mathbb{R}^d$). We remark that the constrain set in (P_K) is defined on the empirical estimate of the risk of $\mathcal{I}_{\boldsymbol{\lambda}}(y)$ and it does not involve the computation of the UCB. Then, (P_K) can be solved with any standard off-the-shelf convex optimization software (e.g., CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018), MOSEK (ApS, 2019)).

[4]We defer proofs to Appendix A.3.

Algorithm 1 K-RCPS

Input: risk level $\epsilon \geq 0$, failure probability $\delta \in [0, 1]$, calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples such that $n = n_{\text{opt}} + n_{\text{RCPS}}$, membership function $\mathcal{M}$, family of set-valued predictors $\{\mathcal{I}_{\boldsymbol{\lambda}}(y) = [\hat{l}_j - \lambda_j, u_j + \lambda_j]\}_{\boldsymbol{\lambda} \in \Lambda^d}$, initial (large) value $\beta_{\max}$, stepsize $\mathrm{d}\beta > 0$.
Split $\mathcal{S}_{\text{cal}}$ into $\mathcal{S}_{\text{opt}}, \mathcal{S}_{\text{RCPS}}$
$M \leftarrow \mathcal{M}(\mathcal{S}_{\text{opt}})$
$\widetilde{\boldsymbol{\lambda}}_K \leftarrow \texttt{SOLVE-PK}(\mathcal{S}_{\text{opt}}, M)$
$\boldsymbol{\lambda} \leftarrow M\widetilde{\boldsymbol{\lambda}}_K + \beta_{\max}\mathbf{1}$
$\hat{R}^{01+}(\boldsymbol{\lambda}) \leftarrow 0$
while $\hat{R}^{01+}(\boldsymbol{\lambda}) \leq \epsilon$ **do**
  $\boldsymbol{\lambda}_{\text{prev}} \leftarrow \boldsymbol{\lambda}$
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} - (\mathrm{d}\beta)\mathbf{1}$
  $\boldsymbol{\lambda} \leftarrow [\boldsymbol{\lambda}]_+$
  $\hat{R}^{01}(\boldsymbol{\lambda}) \leftarrow 1/n_{\text{RCPS}} \cdot \sum_{(x_i, y_i) \in \mathcal{S}_{\text{RCPS}}} \ell^{01}(x_i, \mathcal{I}_{\boldsymbol{\lambda}}(y_i))$
  $\hat{R}^{01+}(\boldsymbol{\lambda}) \leftarrow \text{UCB}(n_{\text{RCPS}}, \delta, \hat{R}^{01}(\boldsymbol{\lambda}))$
end while
$\hat{\boldsymbol{\lambda}}_K \leftarrow \boldsymbol{\lambda}_{\text{prev}}$
return $\hat{\boldsymbol{\lambda}}_K$

Our novel conformal risk control procedure, K-RCPS, finds a vector $\hat{\boldsymbol{\lambda}}_K$ that approximates a solution to the nonconvex optimization problem (P_d) via a two step procedure:

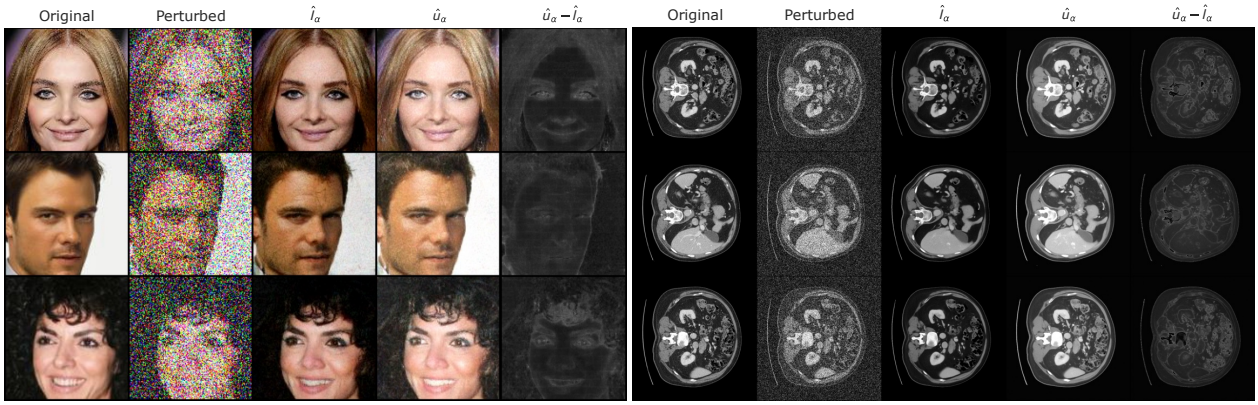
1. First obtaining the optimal solution $\widetilde{\boldsymbol{\lambda}}_K$ to a user-defined (P_K) problem, and then
2. Choosing $\hat{\beta} \in \Lambda$ such that

$$\hat{\beta} = \inf\{\beta \in \Lambda : \; \hat{R}^{01+}(M\widetilde{\boldsymbol{\lambda}}_K + \beta'\mathbf{1}) < \epsilon, \; \forall \beta' \geq \beta\}$$

and return $\hat{\boldsymbol{\lambda}}_K = M\widetilde{\boldsymbol{\lambda}}_K + \hat{\beta}\mathbf{1}$.

Intuitively, the K-RCPS algorithm is equivalent to performing the original RCPS procedure along the line $M\widetilde{\boldsymbol{\lambda}}_K + \beta\mathbf{1}$ parametrized by β. We remark that—as noted in Theorem 3.3—any choice of $\boldsymbol{\eta} \geq 0$ provides a valid direction along which to perform the RCPS procedure. Here, we choose $\mathbf{1}$ because it is precisely the gradient of the objective function. Future work entails devising more sophisticated algorithms to approximate the solution of (P_d).

Algorithm 1 implements the K-RCPS procedure for any calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$, any general family of set-valued predictors of the form $\{\mathcal{I}_{\boldsymbol{\lambda}} = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j]\}_{\boldsymbol{\lambda} \in \Lambda^d}$, any membership function $\mathcal{M} : \; \{\mathcal{X} \times \mathcal{Y}\}^n \to \{0, 1\}^{d \times K}$, and a general $\text{UCB}(n, \delta, \hat{R}(\boldsymbol{\lambda}))$ that accepts a number of



(a) CelebA dataset, $\sigma_0^2 = 1.0,\ \alpha = 0.10$.

(b) AbdomenCT-1K dataset, $\sigma_0^2 = 0.4,\ \alpha = 0.20$.

Figure 3: Calibrated quantiles $\mathcal{I}^\alpha(y)$ computed on 128 samples from $F(y)$ for noisy inputs y with noise level σ_0^2. The difference $\hat{u}_\alpha - \hat{l}_\alpha$ represents intervals sizes (i.e., larger intervals indicate larger uncertainty).

samples n, a failure probability δ, and an empirical risk $\hat{R}(\boldsymbol{\lambda})$ and that returns a pointwise upper confidence bound $\hat{R}^+(\boldsymbol{\lambda})$ that satisfies Equation (6). We remark that, following the *split fixed sequence testing* framework introduced in Angelopoulos et al. (2021) and applied in previous work (Laufer-Goldshtein et al., 2022), the membership matrix and its optimization problem (P_K) are computed on a subset $\mathcal{S}_{\text{opt}}$ of the calibration set $\mathcal{S}_{\text{cal}}$, such that the direction $M\widetilde{\boldsymbol{\lambda}}_K + \beta\mathbf{1}$ along which to perform the RCPS procedure is chosen *before* looking at the data $\mathcal{S}_{\text{RCPS}} = \mathcal{S}_{\text{cal}} \setminus \mathcal{S}_{\text{opt}}$. We note that K-RCPS allows for some of the entries in $\hat{\boldsymbol{\lambda}}_K$ to be set to 0, which preserves the original intervals such that—if they are obtained as described in Section 3.1—they still provide entrywise coverage of future samples at the desired level α.

We now move onto showcasing the advantage of K-RCPS in terms on mean interval length on two real-world high dimensional denoising problems: one on natural images of faces as well as on CT scans of the abdomen.

4. Experiments

As a reminder, the methodological contribution of this paper is two-fold: (i) we propose to use the calibrated quantiles $\mathcal{I}^\alpha(y)$ as a statistically valid notion of uncertainty for diffusion models, and (ii) we introduce the K-RCPS procedure to guarantee high-dimensional risk control. Although it is natural to use the two in conjunction, we remark that the K-RCPS procedure is agnostic of the notion of uncertainty and it can be applied to any nested family of set predictors $\{\mathcal{I}_{\boldsymbol{\lambda}}(y)\}_{\boldsymbol{\lambda}\in\Lambda^d}$ that satisfy the additive parametrization in Equation (13). Therefore, we compare K-RCPS with the original RCPS algorithm on several baseline notions of uncertainty: quantile regression (Angelopoulos et al., 2022b), MC-Dropout (Gal & Ghahramani, 2016), N-Con$ffusion$ (Horwitz & Hoshen, 2022), and naive (i.e., not calibrated) quantiles. We focus on denoising problems where $y = x + v_0$ with $v_0 \sim \mathcal{N}(0, \sigma_0^2)$, on two imaging datasets: the CelebA dataset (Liu et al., 2018) and the AbdomenCT-1K dataset (Ma et al., 2021). In particular—for each dataset—we train:

- A time-conditional score network $s(\tilde{x}, t) \approx \nabla_x \log p_t(\tilde{x})$ following Song et al. (2020b) to sample from the posterior distribution $p(x|y)$ as described in Section 2.1, and
- a time-conditional image regressor $f : \mathcal{Y} \times \mathbb{R} \to \mathcal{X}^3$ following Angelopoulos et al. (2022b) such that $f(y, t) = (\hat{q}_{\alpha/2}, \hat{x}, \hat{q}_{1-\alpha/2})$, where $\hat{x} \approx \mathbb{E}[x \mid y]$ minimizes the MSE loss between the noisy observation y and the ground truth x, and $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantile regressors of x, respectively (Koenker & Bassett Jr, 1978; Romano et al., 2019; Angelopoulos et al., 2022b).

Both models are composed of the same NCSN++ backbone (Song et al., 2020b) with dropout $p = 0.10$ for a fair comparison. We then fine-tune the original score network $s(\tilde{x}, t)$ according to the N-Con$ffusion$ algorithm proposed by Horwitz & Hoshen (2022) such that—similarly to the image regressor f—the resulting time-conditional predictor $\tilde{s}(y, t) = (\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2})$ estimates the $\alpha/2$ and $1-\alpha/2$ quantile regressors of x. Finally, in order to compare with MC-Dropout, we activate the dropout layers in the image regressor f at inference time, and estimate the mean $\bar{x}$ and standard deviation $\hat{\sigma}$ over 128 samples $\hat{x}_1, \ldots, \hat{x}_{128}$. To summarize, we compare K-RCPS and RCPS on the following families of nested set predictors:

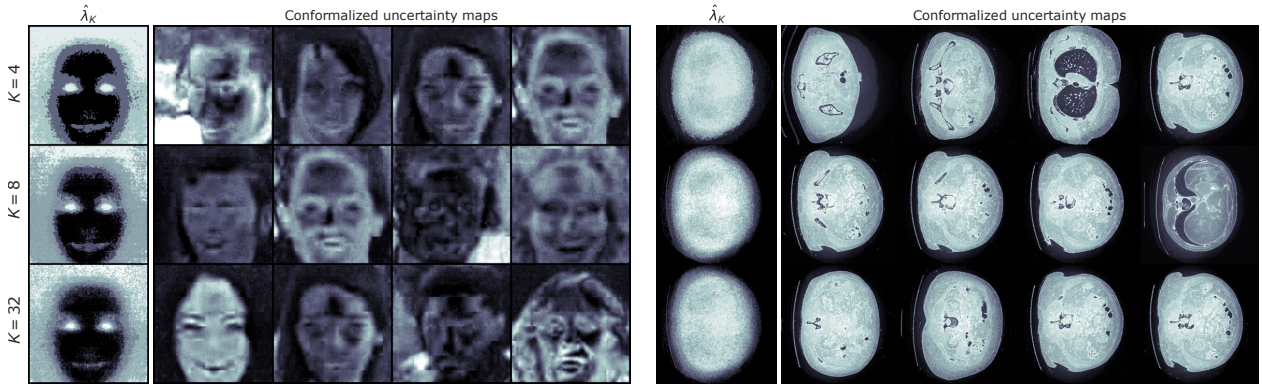
Quantile Regression (QR)

$$\mathcal{I}_{\boldsymbol{\lambda},\text{QR}}(y)_j = [\hat{x}_j - \lambda_j(\hat{q}_{\alpha/2})_j, \hat{x}_j + \lambda_j(\hat{q}_{1-\alpha/2})_j], \quad (17)$$

where $f(y, t) = (\hat{q}_{\alpha/2}, \hat{x}, \hat{q}_{1-\alpha/2})$.

MC-Dropout

$$\mathcal{I}_{\boldsymbol{\lambda},\text{MC-Dropout}}(y)_j = [\bar{x}_j - \lambda_j\hat{\sigma}_j, \bar{x}_j + \lambda_j\hat{\sigma}_j], \quad (18)$$



(a) CelebA dataset ($\alpha = 0.10$, $\epsilon = 0.10$).

(b) AbdomenCT-1K dataset ($\alpha = 0.20$, $\epsilon = 0.05$).

Figure 4: Example optimal $\hat{\boldsymbol{\lambda}}_K$ for $K \in \{4, 8, 32\}$, $n_{\text{opt}} = 256$, and $d_{\text{opt}} = 100$ with respective conformalized uncertainty maps $\mathcal{I}^{\alpha}_{\hat{\boldsymbol{\lambda}}}(y) = [\hat{l}_{j,\alpha} - (\hat{\lambda}_K)_j, \hat{u}_{j,\alpha} + (\hat{\lambda}_K)_j]$. With probability at least 90% no more than ϵ portion of the ground truth pixels will fall outside of $\mathcal{I}^{\alpha}_{\hat{\boldsymbol{\lambda}}}$ on future, unseen samples.

where $\bar{x}, \hat{\sigma}$ are the sample mean and standard deviation over 128 samples $\hat{x}_1, \ldots, \hat{x}_{128}$ obtained by activating the dropout layers in the image regressor f.

N-Con$ffusion$[5] We compare two different parametrizations—multiplicative and additive:

$$\mathcal{I}^{\text{multiplicative}}_{\boldsymbol{\lambda},\text{Conffusion}}(y)_j = \left[\frac{(\hat{q}_{\alpha/2})_j}{\lambda_j}, \lambda_j(\hat{q}_{1-\alpha/2})_j\right] \tag{19}$$

and

$$\mathcal{I}^{\text{additive}}_{\boldsymbol{\lambda},\text{Conffusion}}(y)_j = \left[(\hat{q}_{\alpha/2})_j - \lambda_j, (\hat{q}_{1-\alpha/2})_j + \lambda_j\right], \tag{20}$$

where $\tilde{s}(y, t) = (\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2})$ is the fine-tuned score network by means of quantile regression on 1000 additional samples.

Naive quantiles

$$\mathcal{I}_{\boldsymbol{\lambda},\text{naive}}(y)_j = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j], \tag{21}$$

where $\hat{l}, \hat{u}$ are the naive (i.e., not calibrated) $\alpha/2$ and $1-\alpha/2$ entrywise empirical quantiles computed on 128 samples from the diffusion model.

Calibrated quantiles

$$\mathcal{I}^{\alpha}_{\boldsymbol{\lambda}}(y)_j = [\hat{l}_{j,\alpha} - \lambda_j, \hat{u}_{j,\alpha} + \lambda_j] \tag{22}$$

where $\hat{l}_\alpha, \hat{u}_\alpha$ are the entrywise calibrated quantiles computed on 128 samples from the diffusion model as described in Equation (9) (see Figure 3 for some examples).

We include further details on the datasets, the models, and the training and sampling procedures in Appendix D. The implementation of K-RCPS with all code and data necessary to reproduce the experiments is available at `https://github.com/Sulam-Group/k-rcps`.

[5]A detailed discussion of Con$ffusion$ and the contributions of this paper is included in Appendix E.

We compare all models and calibration procedures on 20 random draws of calibration and validation sets $\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{val}}$ of length n_{cal} and n_{val}, respectively. We remark that for the K-RCPS procedure, n_{opt} samples from $\mathcal{S}_{\text{cal}}$ will be used to solve the optimization problem (P_K). It follows that for a fixed n_{cal}, the concentration inequality used in the K-RCPS procedure will be looser compared to the one in the RCPS algorithm. We will show that there remains a clear benefit of using the K-RCPS algorithm in terms of mean interval length given the same amount of calibration data available (i.e., even while the concentration bound becomes looser). In these experiments, we construct the membership matrix M by assigning each feature $j \in [d]$ to the respective k^{th}, $k = 1, \ldots, K$ quantile of the entrywise empirical estimate of the risk on $\mathcal{S}_{\text{opt}}$. Furthermore, even though (P_K) is low-dimensional (i.e., $K \ll d$), the number of constraints grows as dn_{opt}, which quickly makes the computation of $\widetilde{\boldsymbol{\lambda}}_K$ inefficient and time-consuming (e.g., for the AbdomenCT-1K dataset, $dn_{\text{opt}} \sim 10^8$ when $n_{\text{opt}} = 128$, a mild number of samples to optimize over). In practice, we randomly subsample a small number of features $d_{\text{opt}} \ll d$ stratified by membership, which drastically speeds up computation. In the following experiments, we set $d_{\text{opt}} \in \{50, 100\}$ and use $K \in \{4, 8, 32\}$, which reduces the runtime of solving the optimization problem to less than a second for both datasets. Finally, we pick γ that minimizes the objective function over 16 values equally spaced in $[0.3, 0.7]$. The choice of these heuristics makes the runtime of K-RCPS comparable to that of RCPS, with a small overhead to solve the reduced (P_K) problem (potentially multiple times to optimize γ).

Figure F.1 shows that all combinations of notion of uncertainty and calibration procedures control the risk, as promised. In particular, we set $\delta = 0.10$ for both datasets, and $\epsilon = 0.10, 0.05$ for the CelebA and AbdomenCT-1K dataset, respectively. We repeat all calibration procedures over 20 random samples of $\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{val}}$, with $n_{\text{val}} = 128$, and

Table 1: Comparison of all notions of uncertainty with RCPS and K-RCPS in terms of guarantees provided and mean interval length over 20 independent draws of $\mathcal{S}_{\text{cal}}$. We refer the reader to Appendix E for a detailed discussion of the comparison with the Con$ffusion$ framework.

UNCERTAINTY	DIFFUSION MODEL?	ENTRYWISE COVERAGE?	RISK CONTROL?	CALIBRATION PROCEDURE	MEAN INTERVAL LENGTH CELEBA	MEAN INTERVAL LENGTH ABDOMENCT-1K
QR	✗	✗	✓	RCPS	0.4843 ± 0.0121	0.2943 ± 0.0060
MC-DROPOUT	✗	✗	✓	RCPS	0.6314 ± 0.0109	0.2810 ± 0.0013
N-CON$ffusion$						
— MULTIPLICATIVE	✗	✗	✓	RCPS	0.6949 ± 0.0084	0.1126 ± 0.0020
— ADDITIVE	✗	✗	✓	RCPS	0.3314 ± 0.0040	0.1164 ± 0.0024
— ADDITIVE	✗	✗	✓	K-RCPS	0.3131 ± 0.0056	0.1136 ± 0.0019
NAIVE QUANTILES	✓	✗	✓	RCPS	0.2688 ± 0.0068	0.1518 ± 0.0016
NAIVE QUANTILES	✓	✗	✓	K-RCPS	0.2523 ± 0.0052	0.1374 ± 0.0019
CALIBRATED QUANTILES	✓	✓	✓	RCPS	0.2762 ± 0.0059	0.1506 ± 0.0014
CALIBRATED QUANTILES	✓	✓	✓	K-RCPS	0.2644 ± 0.0067	0.1369 ± 0.0016

$n_{\text{cal}} = 640$ or $n_{\text{cal}} = 512$ for the CelebA or AbdomenCT-1K dataset, respectively. Figure 4 showcases some example $\hat{\boldsymbol{\lambda}}_K$'s obtained by running the K-RCPS procedure with $K = 4, 8$, and 32 quantiles alongside their respective conformalized uncertainty maps from $\mathcal{S}_{\text{val}}$. We can appreciate how for both datasets, $\hat{\boldsymbol{\lambda}}_K$ captures information about the structure of the data distribution (e.g., eyes and lips for the CelebA dataset, and the position of lungs and the heart for the AbdomenCT-1K dataset). Finally, we compare all baselines and calibration procedures in terms of the guarantees each of them provide and their mean interval length. In particular, we report whether each notion of uncertainty provides guarantees over a diffusion model or not. Note that naive and calibrated quantiles are the only notions of uncertainty that precisely provide guarantees on the samples from a diffusion model. Furthermore, calibrated quantiles are the only method that ensures entrywise coverage on future samples on the same noisy observation. For K-RCPS, we perform a grid search over $n_{\text{opt}} \in \{128, 256\}$, $d_{\text{opt}} \in \{50, 100\}$, and $K \in \{4, 8, 32\}$, and we report the optimal results in Table 1. For both datasets, K-RCPS provides the tightest intervals among methods that provide both entrywise coverage and risk control for diffusion models. When relaxing the constraint of entrywise coverage, K-RCPS still provides the tightest intervals. Across the uncertainty quantification methods that do not relate to a diffusion model, we found that K-RCPS with naive sampling provides better results on the CelebA dataset. For the AbdomenCT-1K dataset, N-Con$ffusion$ with multiplicative parametrization and RCPS provides slightly shorter intervals compared to Con$ffusion$ with additive parametrization and K-RCPS. However, we stress that the intervals provided by N-Con$ffusion$ are computed for a fine-tuned model that is different from the original diffusion model, and thus provide no guarantees over the samples of the original model. We found that N-Con$ffusion$ with multiplicative parametrization underperforms on the CelebA dataset because the lower bounds do not decay fast enough, and the loss is concentrated on features whose $(\hat{q}_{\alpha/2})_j/\lambda_j > x_j$.

5. Conclusions

Diffusion models represent huge potential for sampling in inverse problems, alas how to devise precise guarantees on uncertainty has remained open. We have provided *(i)* calibrated intervals that guarantee coverage of future samples generated by diffusion models, *(ii)* shown how to extend RCPS to K-RCPS, allowing for greater flexibility by conformalizing in higher dimensions by means of a convex surrogate problem. Yet, our results are general and hold for any data distribution and any sampling procedure—diffusion models or otherwise. When combined, these two contributions provide state of the art uncertainty quantification by controlling risk with minimal mean interval length. Our contributions open the door to a variety of new problems. While we have focused on denoising problems, the application of these tools for other, more challenging restoration tasks is almost direct since no distributional assumptions are employed. The variety of diffusion models for other conditional-sampling problem can readily be applied here too (Yang et al., 2022; Croitoru et al., 2022). Lastly—and differently from other works that explore controlling multiple risks (Laufer-Goldshtein et al., 2022)—ours is the first approach to provide multi-dimensional control of one risk for conformal prediction, and likely improvements to our optimization schemes could be possible. More generally, we envision our tools to contribute to the responsible use of machine learning in modern settings.

Acknowledgments

The authors sincerely thank Anastasios Angelopoulos for insightful discussions in early stages of this work, as well as the anonymous reviewers who have helped improve our paper.

References

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022a.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022b.

ApS, M. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019. URL http://docs.mosek.com/9.0/toolbox/index.html.

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

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

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Bentkus, V. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Carlier, G., Chernozhukov, V., and Galichon, A. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Carlier, G., Chernozhukov, V., De Bie, G., and Galichon, A. Vector quantile regression and optimal transport, from theory to numerics. *Empirical Economics*, pp. 1–28, 2020.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Horwitz, E. and Hoshen, Y. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Ingraham, J., Baranov, M., Costello, Z., Frappier, V., Ismail, A., Tie, S., Wang, W., Xue, V., Obermeyer, F., Beam, A., et al. Illuminating protein space with a programmable generative model. *bioRxiv*, 2022.

Kadkhodaie, Z. and Simoncelli, E. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021.

Karatzas, I., Karatzas, I., Shreve, S., and Shreve, S. E. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Kawar, B., Vaksman, G., and Elad, M. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021a.

Kawar, B., Vaksman, G., and Elad, M. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1866–1875, 2021b.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Kutiel, G., Cohen, R., Elad, M., and Freedman, D. What's behind the mask: Estimating uncertainty in image-to-image problems. *arXiv preprint arXiv:2211.15211*, 2022.

Laufer-Goldshtein, B., Fisch, A., Barzilay, R., and Jaakkola, T. Efficiently controlling multiple risks with pareto testing. *arXiv preprint arXiv:2210.07913*, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (1):71–96, 2014.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Metzer, G., Richardson, E., Patashnik, O., Giryes, R., and Cohen-Or, D. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.

Pang, T., Xu, K., Li, C., Song, Y., Ermon, S., and Zhu, J. Efficient learning of generative models via finite-difference score matching. *Advances in Neural Information Processing Systems*, 33:19175–19188, 2020.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

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

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Rosenberg, A. A., Vedula, S., Romano, Y., and Bronstein, A. M. Fast nonlinear vector quantile regression. *arXiv preprint arXiv:2205.14977*, 2022.

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

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Torem, N., Ronen, R., Schechner, Y. Y., and Elad, M. Towards a most probable recovery in optical imaging. *arXiv preprint arXiv:2212.03235*, 2022.

Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022.

Xie, Y. and Li, Q. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. *arXiv preprint arXiv:2203.03623*, 2022.

Xu, J., Wang, X., Cheng, W., Cao, Y.-P., Shan, Y., Qie, X., and Gao, S. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., and Kreis, K. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022.

A. Proofs

In this section, we include the proofs for the results presented in this paper. Herein, denote $\mathcal{I}: \mathcal{Y} \to \mathcal{X}'$ a set-valued predictor from $\mathcal{Y} \subset \mathbb{R}^d$ into a space of subsets $\mathcal{X}' \subseteq 2^{\mathcal{X}}$ for $\mathcal{X} \subset \mathbb{R}^d$.

A.1. Proof of Lemma 3.2

Let $F: \mathcal{Y} \to \mathcal{X}$ be a stochastic sampling procedure from $\mathcal{Y}$ to $\mathcal{X}$ such that for a fixed $y \in \mathcal{Y}$, $F(y)$ is a random vector with unknown distribution $\mathcal{Q}_y$. We show that for a desired miscoverage level $\alpha \in [0,1]$, the entrywise calibrated empirical quantiles $\mathcal{I}^\alpha(y)_j = [\hat{l}_{j,\alpha}, \hat{u}_{j,\alpha}]$ defined in Equation (9) provide entrywise coverage as in Definition 3.1. That is, for each $j \in [d] \coloneqq \{1, \dots, d\}$

$$\mathbb{P}[F(y)_j \in \mathcal{I}^\alpha(y)_j] \geq 1 - \alpha. \tag{23}$$

Proof. The proof is a variation of the classical split conformal prediction coverage guarantee (see Angelopoulos & Bates (2021), Theorem D.1). Let $F_1, \dots, F_m, F_{m+1}$ be $m+1$ i.i.d. samples from $F(y)$. For a desired miscoverage level $\alpha \in [0,1]$ and for each $j \in [d]$ denote

$$\hat{l}_{j,\alpha} = \inf\left\{l: \frac{|\{k: (F_k)_j \leq l\}|}{m} \geq \frac{\lfloor (m+1)\alpha/2 \rfloor}{m}\right\} \tag{24}$$

and

$$\hat{u}_{j,\alpha} = \inf\left\{u: \frac{|\{k: (F_k)_j \leq u\}|}{m} \geq \frac{\lceil (m+1)(1-\alpha/2) \rceil}{m}\right\} \tag{25}$$

the $\lfloor (m+1)\alpha/2 \rfloor / m$ and $\lceil (m+1)(1-\alpha/2) \rceil / m$ entrywise calibrated empirical quantiles of $F_1, \dots, F_m$. Assume that for each $j \in [d]$, the first m samples are ordered in ascending order, i.e. $(F_1)_j < \dots < (F_m)_j$ such that

$$\hat{l}_{j,\alpha} = (F_{\lfloor (m+1)\alpha/2 \rfloor})_j \quad \text{and} \quad \hat{u}_{j,\alpha} = (F_{\lceil (m+1)(1-\alpha/2) \rceil})_j. \tag{26}$$

Note that by symmetry of $(F_1)_j, \dots, (F_m)_j$ it follows that $(F_{m+1})_j$ is equally likely to fall between any of the first m samples. That is, for any two indices $m_1 < m_2$

$$\mathbb{P}[(F_{m+1})_j \in [(F_{m_1})_j, (F_{m_2})_j]] = \frac{m_2 - m_1}{m+1}. \tag{27}$$

Instantiating the above equality with $\mathcal{I}^\alpha(y)_j = [\hat{l}_{j,\alpha}, \hat{u}_{j,\alpha}]$ yields

$$\begin{aligned}
\mathbb{P}[(F_{m+1})_j \in \mathcal{I}^\alpha(y)_j] &= \mathbb{P}\left[(F_{m+1})_j \in [\hat{l}_{j,\alpha}, \hat{u}_{j,\alpha}]\right] && (28)\\
&= \mathbb{P}\left[(F_{m+1})_j \in [(F_{\lfloor (m+1)\alpha/2 \rfloor})_j, (F_{\lceil (m+1)(1-\alpha/2) \rceil})_j]\right] && (29)\\
&= \frac{\lceil (m+1)(1-\alpha/2) \rceil - \lfloor (m+1)\alpha/2 \rfloor}{m+1} && (30)\\
&\geq \frac{(m+1)(1-\alpha)}{m+1} = 1 - \alpha && (31)
\end{aligned}$$

which concludes the proof. □

A.2. Proof of Theorem 3.3

Recall that for a calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, a loss function $\ell: \mathcal{X} \times \mathcal{X}' \to \mathbb{R}$, and a family $\{\mathcal{I}_{\boldsymbol{\lambda}}(y)\}_{\boldsymbol{\lambda} \in \Lambda^d}$ of set-valued predictors indexed by a d-dimensional vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d) \in \Lambda^d$, $\Lambda \subset \overline{R} = \mathbb{R} \cup \{\pm\infty\}$

$$R(\boldsymbol{\lambda}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))\right] \quad \text{and} \quad \hat{R}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{(x_i,y_i)\in\mathcal{S}_{\text{cal}}}^{n} \ell(x_i, \mathcal{I}_{\boldsymbol{\lambda}}(y_i)) \tag{32}$$

denote the risk of $\mathcal{I}_{\boldsymbol{\lambda}}(y)$ and its empirical estimate on the calibration set, respectively. Furthermore, let $\hat{R}^+(\boldsymbol{\lambda})$ be a pointwise upper confidence bound (UCB) such that for each fixed $\boldsymbol{\lambda} \in \Lambda^d$ and $\forall \delta \in [0,1]$

$$\mathbb{P}[R(\boldsymbol{\lambda}) \leq \hat{R}^+(\boldsymbol{\lambda})] \geq 1 - \delta \tag{33}$$

as presented in Equation (6). Equivalently to Definition 2.2, for a risk level $\epsilon \geq 0$, we say that $\mathcal{I}_{\boldsymbol{\lambda}}(y)$ is an (ϵ, δ)-RCPS if

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}}[R(\boldsymbol{\lambda}) \leq \epsilon] \geq 1 - \delta. \tag{34}$$

We show that for entrywise monotonically nonincreasing loss functions and for the family of set-valued predictors of the form $\mathcal{I}_{\boldsymbol{\lambda}}(y) = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j]$, for some lower and upper bounds $\hat{l}_j < \hat{u}_j$ that may depend on y, fixed $\boldsymbol{\eta} \geq 0$, if

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda^d} \sum_{j \in [d]} \lambda_j \quad \text{s.t.} \quad \hat{R}^+(\boldsymbol{\lambda} + \beta \boldsymbol{\eta}) < \epsilon, \forall \beta \geq 0 \tag{35}$$

$\mathcal{I}_{\hat{\boldsymbol{\lambda}}}(y)$ is an (ϵ, δ)-RCPS. We start by reminding the following definitions

Definition A.1 (Entrywise monotonically nonincreasing function)**.** A loss function ℓ is entrywise monotonically nonincreasing if for a fixed ground truth x , $\forall j \in [d]$

$$\mathcal{I}(y)_j \subset \mathcal{I}'(y)_j \implies \ell(x, \mathcal{I}'(y)) \leq \ell(x, \mathcal{I}(y)). \tag{36}$$

Definition A.2 (Entrywise nesting property)**.** A family of set predictors $\{\mathcal{I}_{\boldsymbol{\lambda}}(y)\}_{\boldsymbol{\lambda} \in \Lambda^d}$ is entrywise nested if for a fixed observation y, $\forall j \in [d]$

$$\lambda_{j,1} < \lambda_{j,2} \implies \mathcal{I}_{[\lambda_{j,1}, \boldsymbol{\lambda}_{-j}]}(y)_j \subset \mathcal{I}_{[\lambda_{j,2}, \boldsymbol{\lambda}_{-j}]}(y)_j, \tag{37}$$

where $[\lambda_j, \boldsymbol{\lambda}_{-j}]$ is the vector that takes value λ_j in its jth entry and λ_{-j} in its complement $-j := [d] \setminus \{j\}$.

Proof. The proof is a high-dimensional extension of the validity of the original RCPS calibration procedure. Note that the family of set-valued predictors $\{\mathcal{I}_{\boldsymbol{\lambda}}(y) = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j]\}_{\boldsymbol{\lambda} \in \Lambda^d}$ satisfies the entrywise nesting property in Definition A.2. $\forall \hat{\boldsymbol{\lambda}} \in \Lambda^d$, denote $L: \mathbb{R} \to \mathbb{R}$ the one-dimensional function such that

$$L(\beta) = R(\hat{\boldsymbol{\lambda}} + \beta \boldsymbol{\eta}) \quad \text{and} \quad \hat{L}^+(\eta) = \hat{R}^+(\hat{\boldsymbol{\lambda}} + \beta \boldsymbol{\eta}). \tag{38}$$

It follows that L is monotonically nonincreasing because ℓ is entrywise monotonically nonincreasing by assumption, and $\boldsymbol{\eta}$ belongs to the positive orthant. Furthermore, $\mathbb{P}[L(\eta) \leq \hat{L}^+(\eta)] \geq 1 - \delta$ by definition of $\hat{R}^+(\boldsymbol{\lambda})$. Denote

$$\beta^* = \inf\{\beta \in \mathbb{R} : L(\beta) \leq \epsilon\} \tag{39}$$

and assume $R(\hat{\boldsymbol{\lambda}}) = L(0) > \epsilon$. By monotonicity of L it follows that $\beta^* > 0$, and by definition of $\hat{\boldsymbol{\lambda}}$, $\hat{L}^+(\beta^*) = \hat{R}^+(\hat{\boldsymbol{\lambda}} + \beta^* \boldsymbol{\eta}) < \epsilon$. However, since $L(\beta^*) = \epsilon$ and $\mathbb{P}[L(\beta^*) \leq \hat{L}^+(\beta^*)] \geq 1 - \delta$, we conclude that this event happens with probability at most δ. Hence, $\mathbb{P}[R(\hat{\boldsymbol{\lambda}}) \leq \epsilon] \geq 1 - \delta$ and $\mathcal{I}_{\hat{\boldsymbol{\lambda}}}(y)$ is an (ϵ, δ)-RCPS.

Lastly, it is easy to see that $\hat{\boldsymbol{\lambda}}$ minimizes the mean interval length $\bar{I}(\boldsymbol{\lambda})$. Note that

$$\bar{I}(\boldsymbol{\lambda}) = \frac{1}{d} \sum_{j \in [d]} (\hat{u}_j - \hat{l}_j) + 2 \sum_{j \in [d]} \lambda_j, \tag{40}$$

and the statement follows by definition. □

A.3. Proof that ℓ^γ is a convex upper bound to ℓ^{01} (see Equation (16))

Recall that for a family of set-valued predictors $\{\mathcal{I}_{\boldsymbol{\lambda}}(y) = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j]\}_{\boldsymbol{\lambda} \in \Lambda^d}$ indexed by a d-dimensional vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in \Lambda^d$, $\Lambda \subset \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ for some general lower and upper bounds $\hat{l}_j < \hat{u}_j$ that may depend on y, we define $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ to be

$$\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y)) = \frac{1}{d} \sum_{j \in [d]} \left[\frac{2(1+q)}{I(\boldsymbol{\lambda})_j} |x_j - c_j| - q \right]_+, \tag{41}$$

where $q = \gamma/(1-\gamma)$, $\gamma \in [0, 1)$, $I(\boldsymbol{\lambda})_j = \hat{u}_j - \hat{l}_j + 2\lambda_j$, $c_j = (\hat{u}_j + \hat{l}_j)/2$, and $[u]_+ = \max(0, u)$. First, we show that $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ is convex in $\boldsymbol{\lambda}$ for $\boldsymbol{\lambda} \geq 0$.

Proof. Note that $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ is separable in $\boldsymbol{\lambda}$. Hence, it suffices to show that $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ is convex in each entry λ_j. That is, we want to show that

$$\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j = \left[\frac{2(1+q)}{\hat{u}_j - \hat{l}_j + 2\lambda_j}|x_j - c_j| - q\right]_+ \tag{42}$$

is convex in λ_j. Note that:

- The term $1/(\hat{u}_j - \hat{l}_j + 2\lambda_j)$ behaves like $1/\lambda_j$, hence it is convex for $\lambda_j \geq 0 > -(\hat{u}_j - \hat{l}_j)/2$,
- $C = 2(1+q)|x_j - c_j|$ is nonnegative, hence $C \cdot 1/(\hat{u}_j - \hat{l}_j + 2\lambda_j)$ is convex,
- q does not depend on λ_j, hence $C \cdot 1/(\hat{u}_j - \hat{l}_j + 2\lambda_j) - q$ is still convex, and finally
- the positive part $[u]_+ = \max(0, u)$ is a convex function of its argument, hence $[C \cdot 1/(\hat{u}_j - \hat{l}_j + 2\lambda_j) - q]_+$ is convex.

We conclude that $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j$ is convex in each entry $j \in [d]$ for $\lambda_j \geq 0$, hence $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ is convex for $\boldsymbol{\lambda} \geq 0$. □

Note that ℓ^γ is an upper bound to ℓ^{01} by construction. One can see that $\forall \boldsymbol{\lambda},\ \ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y)) \geq \ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$.

Proof. We will now show that $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j \geq \ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j$ entrywise. To this end, we will show that both functions attain the same value (i.e., 1) at the extremes of the intervals (i.e., $\hat{l} - \boldsymbol{\lambda}$ and $\hat{u} + \boldsymbol{\lambda}$), and use the fact that $\ell^\gamma(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j$ is convex nonnegative.

First, note that $\forall \boldsymbol{\lambda}$, it holds that

$$\ell^\gamma(\hat{l} - \boldsymbol{\lambda}, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j = \ell^\gamma(\hat{u} + \boldsymbol{\lambda}, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j = \left[\frac{2(1+q)}{I(\boldsymbol{\lambda})_j} \cdot \left|\frac{I(\boldsymbol{\lambda})_j}{2}\right| - q\right]_+ = 1, \tag{43}$$

$$\ell^{01}(\hat{l} - \boldsymbol{\lambda}, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j = \ell^{01}(\hat{u} + \boldsymbol{\lambda}, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j = 1. \tag{44}$$

Furthermore,

$$\ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))_j = 0,\ \forall x_j \in [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j] \tag{45}$$

by definition. We conclude that ℓ^γ upper bounds ℓ^{01} because ℓ^γ is convex, nonnegative, it intersects ℓ^{01} at $\hat{l} - \boldsymbol{\lambda}$ and $\hat{u} + \boldsymbol{\lambda}$ (for which both losses are equal to 1), and ℓ^{01} is exactly 0 between $\hat{l}_j - \lambda_j$ and $\hat{u}_j + \lambda_j$. □

B. Risk Controlling Prediction Sets (RCPS) (Bates et al., 2021)

In this section, we present the original RCPS procedure presented in Bates et al. (2021). Let $\ell:\ \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$, $\mathcal{Y}' \subseteq 2^{\mathcal{Y}}$ be a monotonically nonincreasing loss function (see Equation (4)) over $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^d$, and let $\{\mathcal{I}_\lambda(y)\}_{\lambda \in \Lambda}$, $\Lambda \subset \overline{\mathbb{R}}$ be a family of set-valued predictors $\mathcal{I}:\ \mathcal{Y} \to \mathcal{X}'$ that satisfies the nesting property in Equation (5). Here,

Algorithm 2 summarizes the original conformal risk control procedure introduced in Bates et al. (2021) for a general bounding function $\text{UCB}(n, \delta, \hat{R}(\lambda))$ that accepts the number of samples in a calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, a failure probability $\delta \in [0, 1]$, the empirical estimate $\hat{R}(\lambda) = 1/n \cdot \sum_{i=1}^n \ell(x_i, \mathcal{I}_\lambda(y_i))$ evaluated on $\mathcal{S}_{\text{cal}}$, and that returns a pointwise upper confidence bound $\hat{R}^+(\lambda)$ that satisfies

$$\mathbb{P}[R(\lambda) \leq \hat{R}^+(\lambda)] \geq 1 - \delta \tag{46}$$

as presented in Equation (6). For example, for losses bounded above by 1, Hoeffding's inequality (Hoeffding, 1994) yields

$$\hat{R}^+(\lambda) = \text{UCB}(n, \delta, \hat{R}(\lambda)) = \hat{R}(\lambda) + \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}. \tag{47}$$

In practice, tighter alternatives exist (see Bates et al. (2021), Section 3.1 for a thorough discussion).

Algorithm 2 (ϵ, δ)-RCPS (see (Angelopoulos et al., 2022b), Algorithm 2)

Input: risk level $\epsilon \geq 0$, failure probability $\delta \in [0,1]$, calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples, monotonically nonincreasing loss function ℓ, family of nested set-valued predictors $\{\mathcal{I}_\lambda\}_{\lambda \in \Lambda}$, $\Lambda \subset \overline{\mathbb{R}}$, initial (large) value $\lambda_{\max}$, stepsize $\mathrm{d}\lambda > 0$.
$\lambda \leftarrow \lambda_{\max}$
$\hat{R}^+(\lambda) \leftarrow 0$
while $\hat{R}^+(\lambda) \leq \epsilon$ **do**
  $\lambda \leftarrow \lambda - \mathrm{d}\lambda$
  $\hat{R}(\lambda) \leftarrow 1/n \cdot \sum_{(x_i,y_i) \in \mathcal{S}_{\text{cal}}} \ell(x_i, \mathcal{I}_\lambda(y_i))$
  $\hat{R}^+(\lambda) \leftarrow \text{UCB}(n, \delta, \hat{R}(\lambda))$
end while
$\hat{\lambda} \leftarrow \lambda + \mathrm{d}\lambda$
return $\hat{\lambda}$

C. Computing $\hat{R}^{01+}(\boldsymbol{\lambda} + \boldsymbol{\eta})$

In this section, we include a detailed analytical example on how one can compute the UCB $\hat{R}^{01+}(\boldsymbol{\lambda} + \boldsymbol{\eta})$ by means of concentration inequalities. Recall that, for a family of nested set predictors $\{\mathcal{I}_{\boldsymbol{\lambda}}(y)\}_{\boldsymbol{\lambda} \in \Lambda^d}$ indexed by a *vector* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, the 01 loss $\ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}}(y))$ counts the number of pixels in x that fall outside of their respective intervals in $\mathcal{I}_{\boldsymbol{\lambda}}(y)$. For example, choosing $\mathcal{I}_{\boldsymbol{\lambda}}(y)_j = [\hat{l}_j - \lambda_j, \hat{u}_j + \lambda_j]$ as proposed in Equation (11), given a vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ yields

$$\ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}+\boldsymbol{\eta}}(y)) = \frac{1}{d} \sum_{j \in [d]} \mathbb{1}\left[x_j \notin [\hat{l}_j - (\lambda_j + \eta_j), \hat{u}_j + (\lambda_j + \eta_j)]\right]. \tag{48}$$

Then, $R^{01}(\boldsymbol{\lambda} + \boldsymbol{\eta}) = \mathbb{E}[\ell^{01}(x, \mathcal{I}_{\boldsymbol{\lambda}+\boldsymbol{\eta}}(y))]$ denotes the risk of $\mathcal{I}_{\boldsymbol{\lambda}+\boldsymbol{\eta}}$ and its empirical estimate over a calibration set $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is

$$\hat{R}^{01}(\boldsymbol{\lambda} + \boldsymbol{\eta}) = \frac{1}{n} \sum_{i \in [n]} \ell^{01}(x_i, \mathcal{I}_{\boldsymbol{\lambda}+\boldsymbol{\eta}}(y_i)). \tag{49}$$

In turn, for a user-specified failure probability $\delta \in [0,1]$, the UCB $R^{01+}(\boldsymbol{\lambda} + \boldsymbol{\eta})$ can be obtained by means of common concentration inequalities. For example, Hoeffding's inequality implies

$$\hat{R}^{01+}(\boldsymbol{\lambda} + \boldsymbol{\eta}) = \hat{R}^{01}(\boldsymbol{\lambda} + \boldsymbol{\eta}) + \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}. \tag{50}$$

D. Experimental Details

In this section, we include further experimental details. All experiments were performed on a private cluster with 8 NVIDIA RTX A5000 with 24 GB of memory.

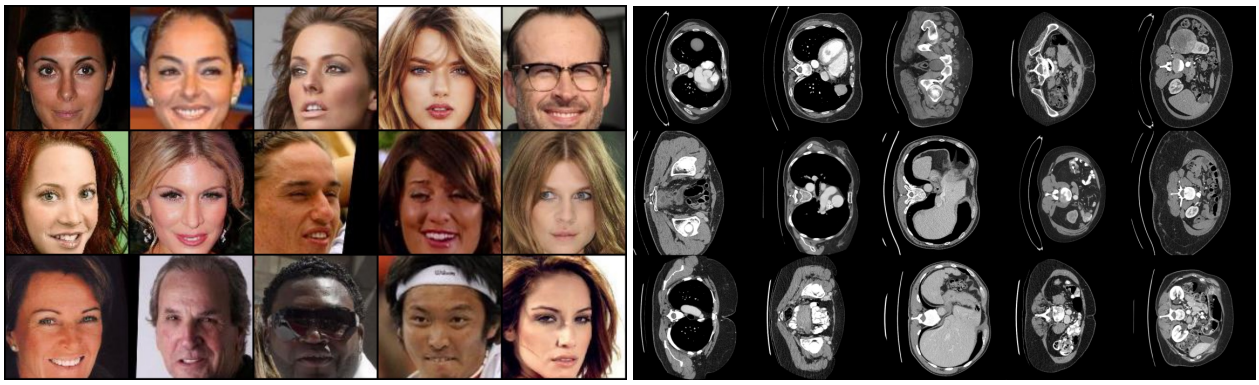
D.1. Datasets

D.1.1. CelebA dataset

The CelebA dataset (Liu et al., 2018) (available at `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`) contains more than 200×10^3, 178×218 pixel images of celebrity faces with several landmark locations and binary attributes (e.g., eyeglasses, bangs, smiling). In this work, we center-crop all images to 128×128 pixels and normalize them between $[-1, 1]$.

D.1.2. AbdomenCT-1K dataset

The AbdomenCT-1K dataset (Ma et al., 2021) (available at `https://github.com/JunMa11/AbdomenCT-1K`) comprises more than 1000 abdominal CT scans (for a total of more than 200×10^3, 512×512 pixel individual images)



(a) CelebA dataset. (b) AbdomenCT-1K dataset.

Figure D.1: Example images.

aggregated from 6 existing datasets. Scans are provided in NIfTI format, so we first convert them to their Hounsfield unit (HU) values and subsequently window them with the standard abdomen setting (WW = 400, WL = 40) such that pixels intensities are normalized in $[0, 1]$ and they represent the same tissue across images.

Figure D.1 shows some example images for both datasets.

D.2. Models

Recall that in this work, we train both:

- A score network $s(\tilde{x}, t) \approx \nabla_x \log p_t(\tilde{x})$ to sample from $p(x|y)$ as introduced in Section 2.1, and
- a modified time-conditional image regressor $f : \mathcal{Y} \times \mathbb{R} \to \mathcal{X}^3$ such that $f(y, t) = (\hat{q}_{\alpha/2}, \hat{x}, \hat{q}_{\alpha/2})$, where $\hat{x} \approx \mathbb{E}[x \mid y]$ and $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ are the entrywise $\alpha/2$ and $1 - \alpha/2$ quantile regressors of x (Koenker & Bassett Jr, 1978; Romano et al., 2019; Angelopoulos et al., 2022b), respectively.

Both models are implemented wit the same U-net-like (Ronneberger et al., 2015) backbone: NCSN++, which was introduced by Song et al. (2020b) (code is available at `https://github.com/yang-song/score_sde`). We use the original NCSN++ configurations presented in Song et al. (2020b) for the CelebA dataset and, for the AbdomenCT-1K dataset, for the FFHQ dataset (available at `https://github.com/NVlabs/ffhq-dataset`) given the larger image size of 512×512. For the image regressor f, we use the original implementation of the quantile regression head used in Angelopoulos et al. (2022b) (available at `https://github.com/aangelopoulos/im2im-uq`) on top of the NCSN++ backbone. This allows us to maintain a time-conditional backbone and extend the original image regressor presented in Angelopoulos et al. (2022b) to all noise levels as the score network.

D.3. Training

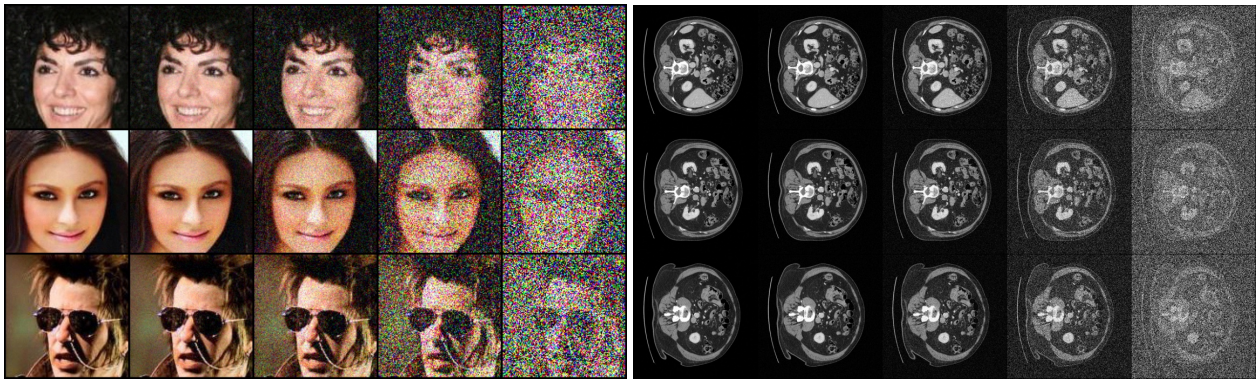
D.3.1. Data Augmentation

For both datasets, we augment the training data by means of random horizontal and vertical flips, and random rotations between $[-\pi/2, \pi/2]$ degrees.

D.3.2. Forward SDE

Recall that in this work, we are interested in solving the classical denoising problem where $y = x + v_0$ is a noisy observation of a ground truth image x perturbed with random Gaussian noise with known variance σ_0^2. As done by previous works (Song & Ermon, 2019; 2020; Song et al., 2021; Kawar et al., 2021a), we model the observation process with a *variance exploding* (VE) forward-time SDE

$$\mathrm{d}x = \sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}\,\mathrm{d}w \quad \text{where} \quad \sigma(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t,\ t \in [0, 1] \tag{51}$$



(a) CelebA dataset. (b) AbdomenCT-1K dataset.

Figure D.2: Example of perturbed images via the forward SDE. The final level of noise is $\sigma^2 = 1$.

such that $\sigma(0) = \sigma_{\min}$ and $\sigma(1) = \sigma_{\max}$. In particular, we set $\sigma_{\min} = 0$ and $\sigma_{\max} = 90$ for the CelebA dataset and $\sigma_{\max} = 1$ for the AbdomenCT-1K dataset. It has been shown (Song et al., 2020b) that the above forward-time SDE is equivalent to the following discrete Markov chain

$$x_t = x_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} z, \; z \sim \mathcal{N}(0, \mathbb{I}), \quad \text{for} \quad t = 1, \dots, N \tag{52}$$

and $\{\sigma_i\}_{i=0}^N$ noise levels. That is, $x_t = x + z_t$, where $z_t \sim (0, \sigma_t^2)$. Figure D.2 shows some example images from both datasets perturbed via the forward SDE described in this section.

D.3.3. Denoising Score-matching

Here, we briefly describe the loss function used to train the time-conditional score network $s(\tilde{x}, t)$. Denote $\theta \in \Theta$ the parametrization of s, then, following Song & Ermon (2019); Song et al. (2020b), we have $s(\tilde{x}, t) = s_{\theta^*}(\tilde{x}, t)$, where

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{t \sim U(0,1)} \left[\xi(t) \mathbb{E}_{x \sim p(x),\, x(t)|x} \left[\| s_\theta(x(t), t) - \nabla_x \log p_t(x(t)|x) \|^2 \right] \right] \tag{53}$$

$$= \arg\min_{\theta \in \Theta} \mathbb{E}_{t \sim U(0,1)} \left[\xi(t) \mathbb{E}_{x \sim p(x),\, x(t)|x} \left[\left\| s_\theta(x(t), t) + \frac{x(t) - x}{\sigma(t)} \right\|^2 \right] \right], \tag{54}$$

where $\xi(t) \propto \sigma^2(t)$ and $U(0, 1)$ is the uniform distribution over $[0, 1]$.

D.3.4. Quantile Regression

Similarly to above, we briefly describe the loss function used to train the modified time-conditional image regressor f. Denote $\theta \in \Theta$ the parametrization of f, and recall that for a desired α quantile and its respective quantile regressor $\hat{q}_\alpha$, the *quantile loss function* (Koenker & Bassett Jr, 1978; Romano et al., 2019; Angelopoulos et al., 2022b) $\ell^\alpha(x, \hat{q}_\alpha)$ is

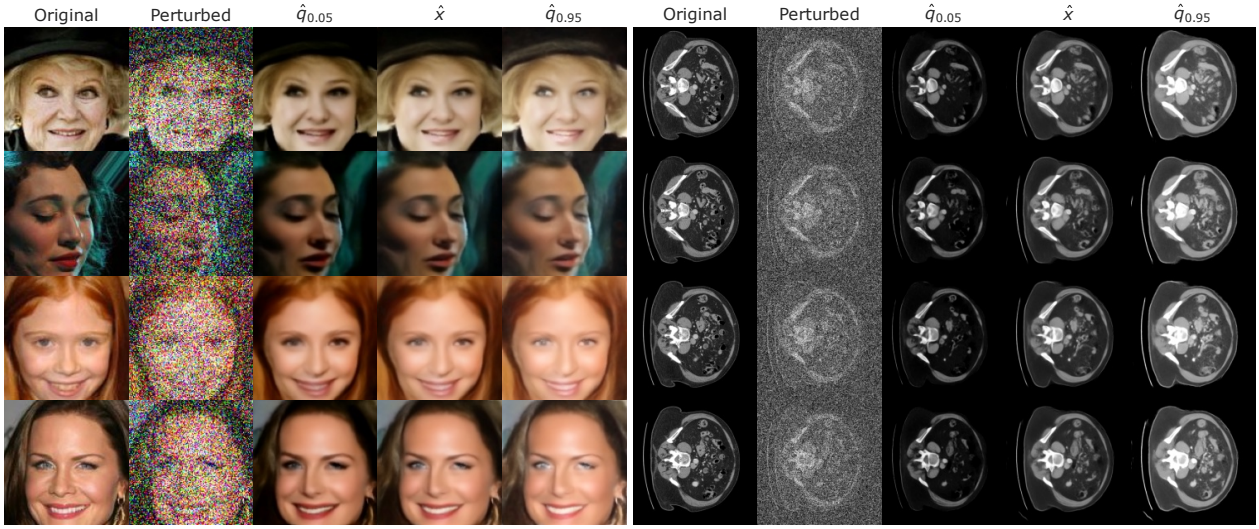
$$\ell^\alpha(x, \hat{q}_\alpha) = \alpha(x - \hat{q}_\alpha) \cdot \mathbb{1}[x > \hat{q}_\alpha] + (1 - \alpha)(\hat{q}_\alpha - x) \cdot \mathbb{1}[x \leq \hat{q}_\alpha]. \tag{55}$$

For the image regressor f, we set $\alpha = 0.10$ and use the original implementation of the quantile regression head from Angelopoulos et al. (2022b) (available at `https://github.com/aangelopoulos/im2im-uq`) which minimizes the multi-loss objective

$$\mathcal{L}(x, f(y, t)) = \ell^{\alpha/2}(x, \hat{q}_{\alpha/2}) + (x - \hat{x})^2 + \ell^{1-\alpha/2}(x, \hat{q}_{1-\alpha/2}) \tag{56}$$

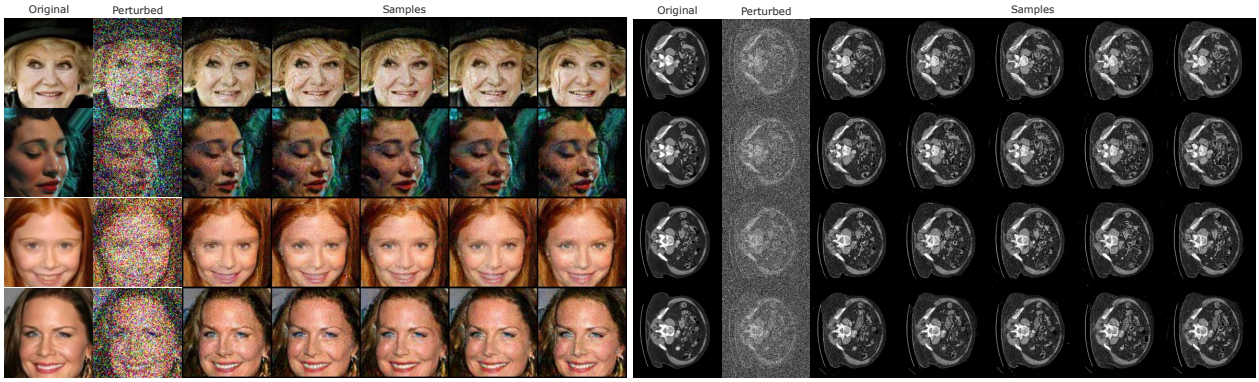
on top of the NCSN++ backbone. This allows us to maintain a time-conditional backbone and extend the original image regressor presented in Angelopoulos et al. (2022b) to all noise levels as the score network. Then, $f(y, t) = f_{\theta^*}(y, t)$ where

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{t \sim U(0,1)} \left[\mathcal{L}(x, f(x(t), t) \right]. \tag{57}$$



(a) CelebA dataset. (b) AbdomenCT-1K dataset.

Figure D.3: Example of results with the modified image regressor.



(a) CelebA dataset. (b) AbdomenCT-1K dataset.

Figure D.4: Example of images sampled via Algorithm 3.

D.4. Sampling

Here, we briefly describe the sampling procedure used in this work to solve the conditional reserve-time SDE

$$\mathrm{d}x = [h(x,t) - g(t)^2 \nabla_x \log p_t(x|y)]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{w} \tag{58}$$

where y is the initial noisy observation with known noise level σ_0^2, and $g(t) = \sqrt{\mathrm{d}[\sigma^2(t)]/\mathrm{d}t},\ h(x,t) = 0$ as in Equation (51). Although several different discretization schemes exist (Song et al., 2020b), we use the classical Euler-Maruyama discretization (Karatzas et al., 1991) since the contribution of this work is not focused on improving existing diffusion models. Recall that for a general Itô process

$$\mathrm{d}x = h(x,t)\,\mathrm{d}t + g(t)\,\mathrm{d}w, \tag{59}$$

its Euler-Maruyama discretization with step Δt is

$$x_{t+1} = x_t + h(x,t)\,\Delta t + g(t)\sqrt{\Delta t}\,z,\ z \sim \mathcal{N}(0, \mathbb{I}). \tag{60}$$

Furthermore, under the forward SDE described in Equation (51), and as shown by previous works (Kawar et al., 2021b;a; Kadkhodaie & Simoncelli, 2021), it holds that

$$\nabla_x \log p_t(x(t)|y) = \nabla_x \log p_t(x(t)) + \frac{y - x(t)}{\sigma_0^2 - \sigma^2(t)}. \tag{61}$$

We conclude that the reverse-time Euler-Maruyama discretization of the conditional SDE above is

$$x_{t-1} = x_t + g_t^2 \left[\nabla_x \log p_t(x_t) + \frac{y - x_t}{\sigma_0^2 - \sigma_t^2} \right] \Delta t + g_t \sqrt{\Delta t}\, z,\ z \sim \mathcal{N}(0, \mathbb{I}). \tag{62}$$

Algorithm 3 implements the sampling procedure for a general score network $s(\tilde{x}, t)$.

Algorithm 3 Denoising Reverse-time SDE

Input: observation y, initial noise level $\sigma_0, \sigma_{\max}, \sigma_{\min}$, number of steps N
$t_0 \leftarrow (\log \sigma_0 - \log \sigma_{\min})/(\log \sigma_{\max} - \log \sigma_{\min})$
$\Delta t \leftarrow 1/N$
$n \leftarrow \lfloor t_0/\Delta t \rfloor$
$x_n \leftarrow y$
for $i = n, \ldots, 1$ **do**
  Draw $z_i \sim \mathcal{N}(0, \mathbb{I})$
  $t_i \leftarrow i\Delta t$
  $\sigma_i \leftarrow \sigma_{\min} \cdot (\sigma_{\max}/\sigma_{\min})^{t_i}$
  $g_i \leftarrow \sigma_i \cdot \sqrt{2 \log(\sigma_{\max}/\sigma_{\min})}$
  $x_{i-1} \leftarrow x_i + g_i^2 [s(x_i, t_i) + (y - x_i)/(\sigma_0^2 - \sigma_i^2)]\, \Delta t + g_i \sqrt{\Delta t}\, z_i$
end for
return x_0

E. Comparison with Con$ffusion$

In this section, we discuss the differences between the Con$ffusion$ framework proposed by Horwitz & Hoshen (2022) and the contributions of this paper. Although similar in spirit and motivation, the two are fundamentally different and address separate and distinct questions. Con$ffusion$ deploys ideas of quantile regression (Koenker & Bassett Jr, 1978) to fine-tune an existing score network and obtain an image regressor equipped with heuristic uncertainty intervals. Such intervals can then be conformalized to provide risk control. On the other hand, our K-RCPS is a novel high-dimensional calibration procedure for any stochastic sampler and any notion of uncertainty (i.e., it is agnostic of the notion of uncertainty), including diffusion models and quantile regression. In this paper, we present K-RCPS for diffusion models given their remarkable performance in solving inverse problems via conditional sampling.

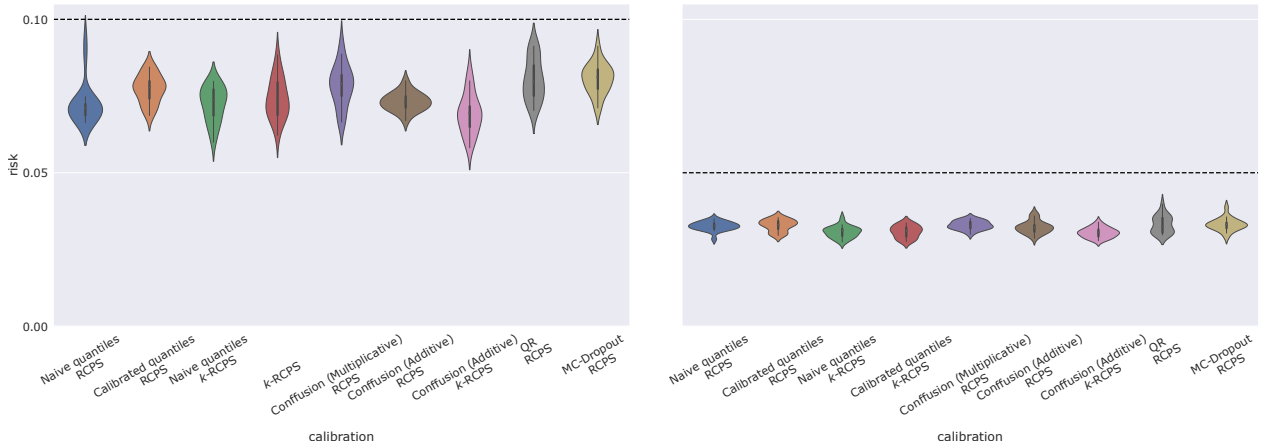
In order to contextualize the choices of parametrization for the families of set predictors in Equations (19) and (20), we highlight a gap between the presentation of Con$ffusion$ in the arXiv paper available at https://arxiv.org/abs/2211.09795v1 and the official code release in https://github.com/eliahuhorwitz/Conffusion. In particular Section 3.2 of Horwitz & Hoshen (2022) introduces the following family of set predictors:

$$\mathcal{I}^{\text{arXiv}}_{\boldsymbol{\lambda},\text{Conffusion}}(y)_j = [\lambda_j(\hat{q}_{\alpha/2})_j, \lambda_j(\hat{q}_{1-\alpha/2})_j], \tag{63}$$

which does not satisfy the nesting property in Equation (5). It is easy to see that—without any further assumptions—as λ_j increases, $\mathcal{I}^{\text{arXiv}}_{\boldsymbol{\lambda},\text{Conffusion}}(y)_j$ does not cover the interval $[0, 1]$ and the 01 loss in Equation (10) is not entrywise monotonically non-increasing. This renders the RCPS procedure (and most calibration procedures) not applicable as presented in the arXiv paper. While this family of set predictors is thus not applicable, we resorted to the authors' GitHub release where such parametrization is different (see https://github.com/eliahuhorwitz/Conffusion/blob/fffe5c946219cf9dead1a1c921a131111e31214e/inpainting_n_conffusion/core/calibration_masked.py#L28). More precisely, the GitHub implementation reflects the multiplicative parametrization presented in Equation (19), which does indeed satisfy the entrywise nesting property. Finally, we additionally propose the additive parametrization in Equation (20) to showcase how K-RCPS is agnostic of the notion of uncertainty and it can be used in conjunction with Con$ffusion$.

F. Figures

This section contains supplementary figures.



(a) CelebA dataset.

(b) AbdomenCT-1K dataset.

Figure F.1: Empirical estimates of risk over 20 random draws of $\mathcal{S}_{\text{cal}}$. All combinations of notion of uncertainty and calibration procedure successfully control risk at level $\epsilon = 0.10, 0.05$ for the CelebA and AbdomenCT-1K dataset, respectively, with probability at least $1 - \delta,\ \delta = 0.10$.