From Uncertain to Safe: Conformal Adaptation of Diffusion Models for Safe PDE Control

Peiyan Hu[2#*], Xiaowei Qian[1#*], Wenhao Deng[1], Rui Wang[3#], Haodong Feng[1], Ruiqi Feng[1], Tao Zhang[1], Long Wei[1],

Yue Wang[4], Zhi-Ming Ma[2], Tailin Wu[1†]

[1] Department of Artificial Intelligence, School of Engineering, Westlake University,

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences,

[3] Fudan University, [4] Zhongguancun Academy

(* equal contributions; [#] intern at Westlake University, [†] corresponding author)

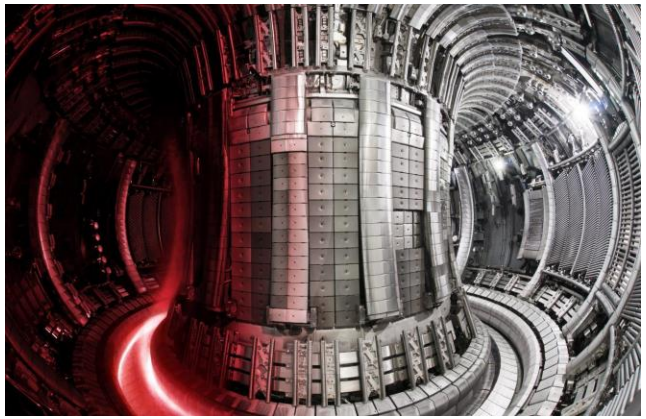
Corresponding to: {hupeiyan,wutailin}@westlake.edu.cn

Introduction

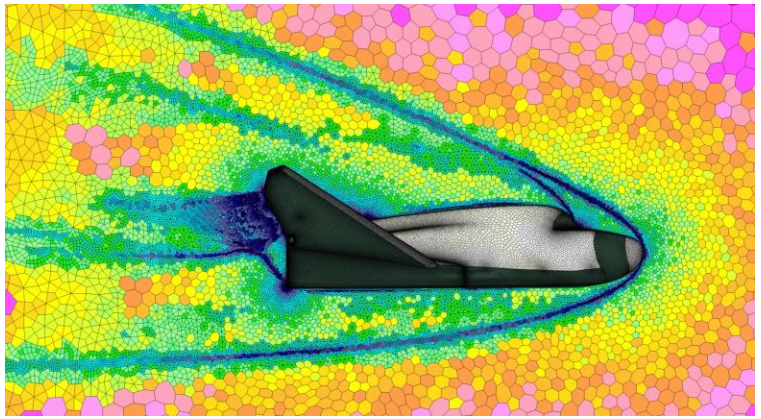
- Safe PDE control task: given a control objective $\mathcal{J}$, find the optimal control signal $\mathbf{w}^*$ while satisfying PDE constraints and constraining the safety score s to stay below the bound s_0 :

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}}\, \mathcal{J}(\boldsymbol{u}, \boldsymbol{w}) \quad \text{s.t.} \quad \boldsymbol{\mathcal{C}}(\boldsymbol{u}, \boldsymbol{w}) = \mathbf{0} \quad \boldsymbol{s}(\boldsymbol{u}) \leq \boldsymbol{s_0}$$

- E.g. How to control external forces on a fluid, to maximize smoke reaching a target exit , under the constraints of fluid dynamics and a hazardous region.



Nuclear Fusion control



Fluid Dynamics

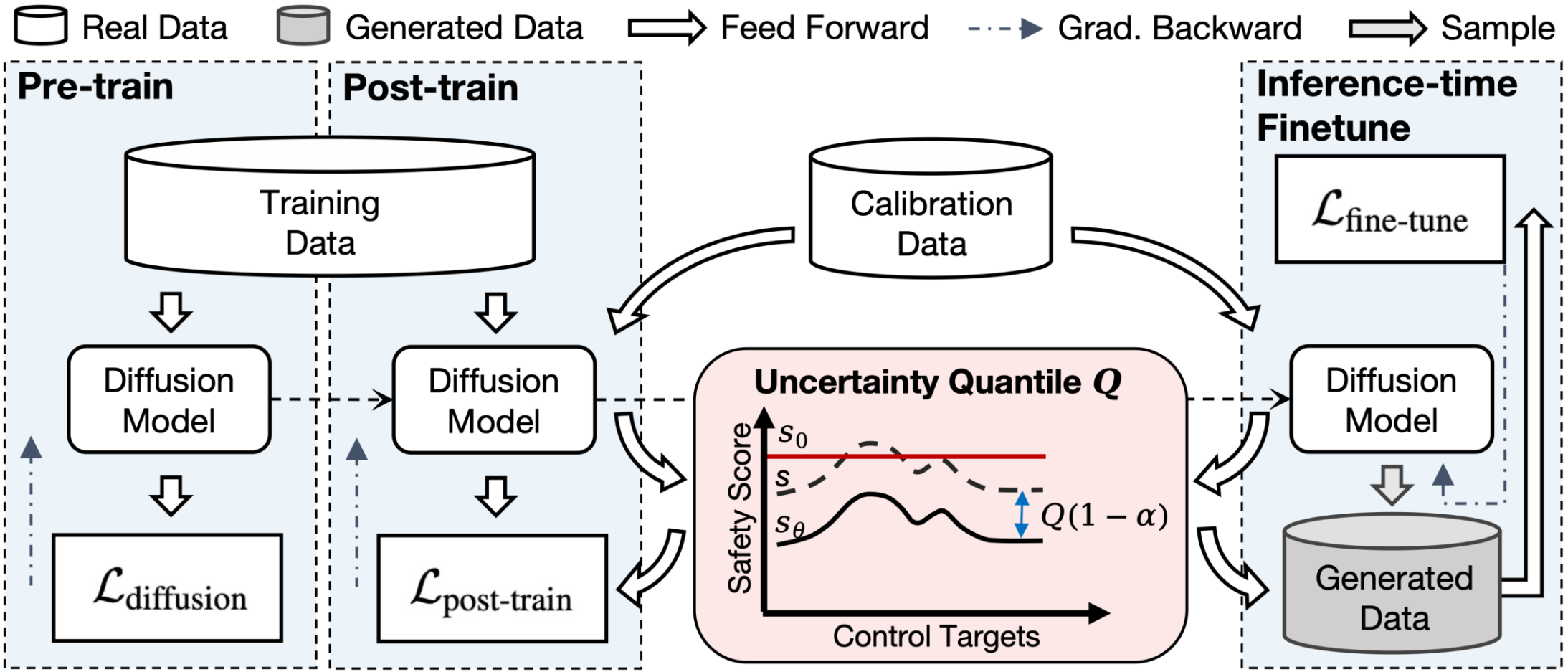
Motivation

- Challenge 1: Suboptimal & Unsafe offline data
  - Models learn from offline data that are filled with suboptimal and unsafe examples
- Challenge 2: Balancing Performance and Safety
  - There's an inherent conflict between optimizing control performance and satisfying safety constraints.
- Our Insight: Quantify and Adapt to Uncertainty
  - We use conformal prediction to quantify this uncertainty. Instead of a single point estimate, we compute a guaranteed safety interval. We then adapt our diffusion model to ensure this entire interval—not just the prediction—stays within the safe boundary.

Preliminary – Conformal Prediction

- Core idea: Use a calibration set to estimate future prediction errors, providing a statistically valid prediction interval with a guaranteed coverage probability of at least $1 - \alpha$
- Calibration Set: Split out from training data, used to estimate the model's prediction errors
- Conformal Scores: For a model prediction $\mu_\theta(X_i)$, a set of error scores $S_i = |\mu_\theta(X_i) - Y_i|$ calculated on the calibration set.
- Significance Level (α): The allowed error rate
- Quantile ($q_{1-\alpha}$): The $(1 - \alpha)$-th quantile of the conformal scores.
- Prediction Interval: For a new point X_{new}, the true value Y_{new} is guaranteed to be in $[\mu_\theta(X_{new}) - q_{1-\alpha}, \mu_\theta(X_{new}) + q_{1-\alpha}]$ with at least $1 - \alpha$ probability.

Method


Real Data
Generated Data
Feed Forward
Grad. Backward
Sample
Pre-train
Post-train
Inference-time Finetune
Training Data
Calibration Data
$\mathcal{L}_{\text{fine-tune}}$
Diffusion Model
Diffusion Model
Diffusion Model
$\mathcal{L}_{\text{diffusion}}$
$\mathcal{L}_{\text{post-train}}$
Uncertainty Quantile Q
Safety Score
s_0
s
s_θ
$Q(1-\alpha)$
Control Targets
Generated Data

Method – Uncertainty Quantification of Diffusion Models

- **Problem:** The standard assumption for conformal prediction doesn't hold. There is a distribution shift between the calibration data and the control sequences generated by the diffusion model during inference.
- **Shifted Score Set:**
  1. Standard Score Set: $\mathcal{S} := \{|s(\mathbf{u}_\theta(\mathbf{w}_i)) - s(\mathbf{u}_i)| : (\mathbf{u}_i, \mathbf{w}_i) \in D_{\text{cal}}\} \cup \{\infty\}$
  2. Re-weight these scores: $\tilde{\mathcal{S}} := \{\omega_{\text{norm}}(\mathbf{u}_i, \mathbf{w}_i)\Delta s_i : \Delta s_i \in \mathcal{S}\}$

  The weight $\omega(\mathrm{u}_i, \mathrm{w}_i)$ estimates the likelihood ratio between the model's target distribution and the calibration distribution
- **Conformal Interval:**

$$CI_\theta(1-\alpha, D_{cal}) := [s(u_\theta(w)) - Q(1-\alpha; \tilde{S}), s(u_\theta(w)) + Q(1-\alpha; \tilde{S})]$$

Method - Post-training with Reweighted Loss

- **Goal:** Steer the pre-trained diffusion model's distribution towards a target distribution that is both safer and more optimal.
- **Uncertainty-Aware Weighting Function:**

$$\mathcal{W}(u, w) = \max[\underbrace{s(u) + Q(1-\alpha; \tilde{S})}_{\text{Upper Bound of CI}} - s_0, 0] + \gamma \mathcal{J}(u, w)$$

  - Penalize unsafe actions and suboptimal objectives. Critically, it penalizes trajectories where the upper bound of the conformal interval exceeds the safety threshold.

- **Reweighted Diffusion Loss :** Modify the standard diffusion training loss by reweighting each sample from the training data

$$\mathcal{L}_{post-train} := \mathbb{E}[e^{-\mathcal{W}(u,w)} ||\epsilon - \epsilon_\theta(\dots)||_2^2]$$

Method – Inference-time Fine-tuning

- **Goal:** For a specific control task at inference time, we further optimize the model to improve safety and performance through an iterative process.
- **Two-Step Iterative Loop:**
  1. Guided Sampling: Generate control sequences using the diffusion model, but guide the sampling process at each denoising step.

$$\mathcal{G}(u, w) = \mathcal{W}(u, w)$$

  2. Fine-tuning: Use the control sequences generated to perform a few steps of gradient descent on the model's parameters θ:

$$\mathcal{L}_{fine-tune} = \sum_{(u_\theta, w_\theta) \in D_{sampled}} \mathcal{W}(u_\theta, w_\theta)$$

Results – New Datasets & Key Findings

- We design three safe control tasks and evaluate our method in them:
  - 1D Burgers' equation
  - 2D incompressible fluid
  - Tokamak fusion reactor

- SafeDiffCon demonstrates superior Safety and control performance:
  - Safety: Across all experiments, SafeDiffCon was the only method that satisfied all safety constraints (0% unsafe trajectories), whereas all classical and deep learning baselines failed on at least one task.
  - Control Performance: While guaranteeing safety, SafeDiffCon also achieved the best control performance among all methods.

Control Results - 1D Burgers' equation

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} = -\mathbf{u} \cdot \dfrac{\partial \mathbf{u}}{\partial \boldsymbol{x}} + \nu \dfrac{\partial^2 \mathbf{u}}{\partial \boldsymbol{x}^2} + \mathbf{w}(t, \boldsymbol{x}), \text{in } [0, T] \times \Omega \\ \mathrm{u}(t, \boldsymbol{x}) = \mathbf{0}, \qquad \text{on } [0, T] \times \partial\Omega \\ \mathrm{u}(0, \boldsymbol{x}) = \mathbf{u_0}(\boldsymbol{x}), \qquad \text{in } \{0\} \times \Omega \end{cases}$$

Control objective: ($u_d(x)$ is target state)

$$J_{actual} = \int |u(T, x) - u_d(x)|^2 \, \mathrm{d}x$$

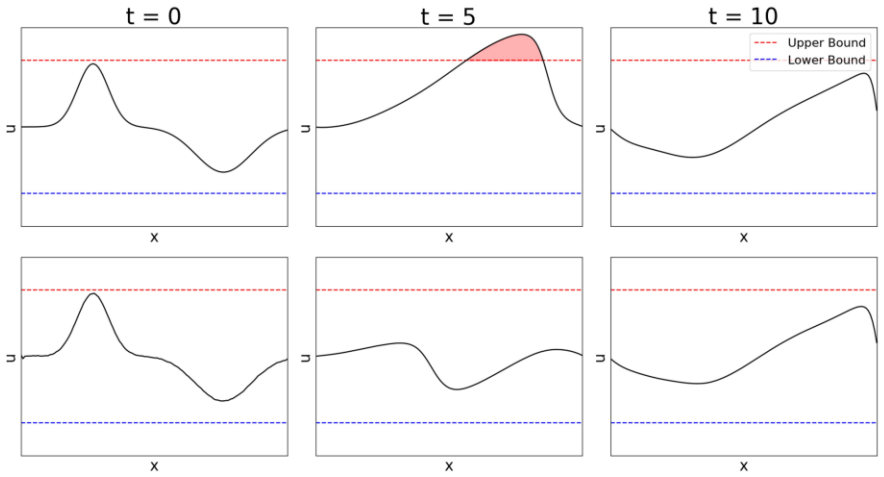
Safety score:

$$s(u) = \sup_{(\mathrm{t},\mathrm{x}) \in [0,T] \times \Omega} \{\mathrm{u}(\mathrm{t}, \mathrm{x})^2\}$$

R_{sample}: unsafe trajectories / total trajectories

R_{time}: unsafe timesteps among all timesteps

R_{point}: unsafe spatial lattice points in all points



Original trajectory

Controlled by SafeDiffCon

Methods	$\mathcal{J} \downarrow$	$\mathcal{R}_{\text{sample}} \downarrow$	$\mathcal{R}_{\text{time}} \downarrow$	$\mathcal{R}_{\text{point}} \downarrow$
BC	0.0001	38%	13%	1.2%
BC-Safe	0.0002	14%	3%	0.2%
PID	0.0968	0%	0%	0.0%
SL-Lag	0.0115	0%	0%	0.0%
MPC-Lag	0.0092	0%	0%	0.0%
CDT	0.0012	16%	3%	0.2%
TREBI	0.0074	0%	0%	0.0%
SafeDiffCon	**0.0016**	0%	0%	0.0%

Control Results - 2D incompressible fluid

Control objective:

the negative ratio of smoke passing through the target bucket located at the center top.

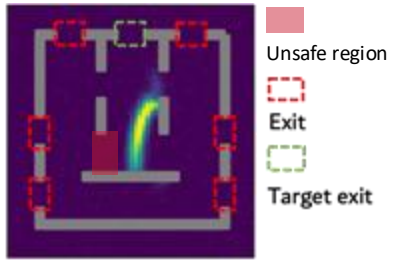
Safety score:

the ratio of smoke entering the unsafe red region.

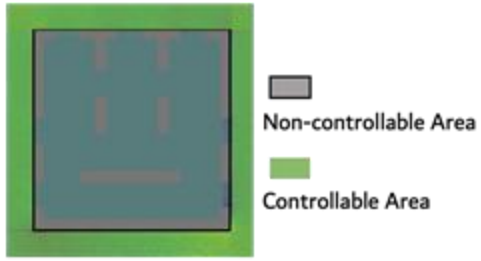
$SVM = \max[s - s_0, 0]$

R: unsafe trajectories

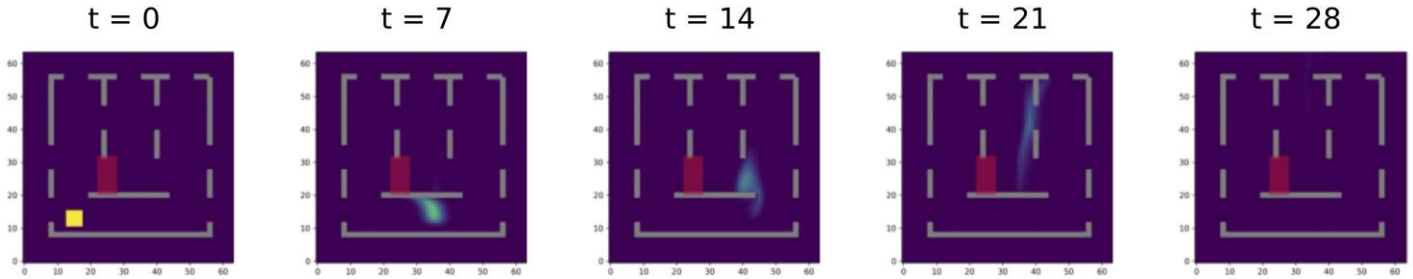
Methods	$\mathcal{J} \downarrow$	$SVM \downarrow$	$\mathcal{R} \downarrow$
BC	-0.7104	0.7156	88%
BC-Safe	-0.2520	0.0330	8%
CDT	-0.7025	0.2519	30%
TREBI	-0.7019	0.0808	18%
SafeDiffCon	**-0.3548**	0.0000	0%



(a) Locations of exits and obstacles



(b) Locations of controllable area



Control Results – Tokamak Fusion Reactor

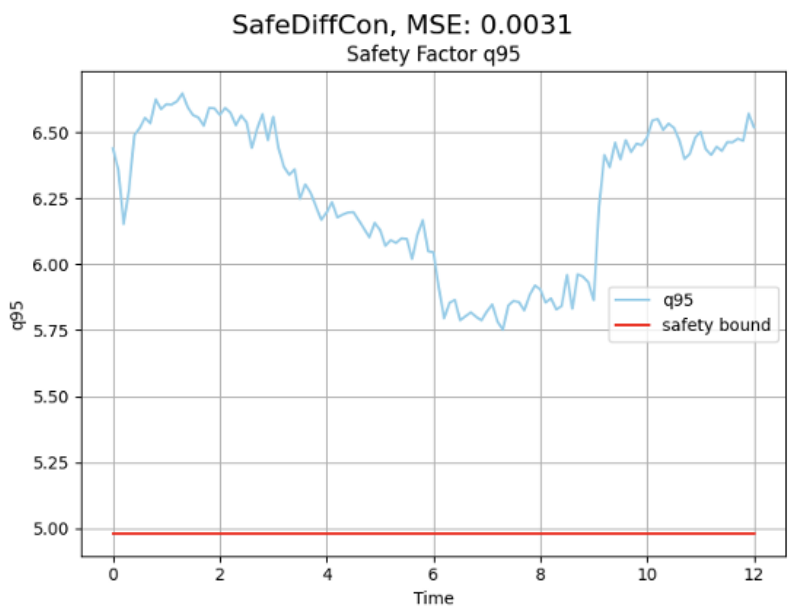
Control objective:

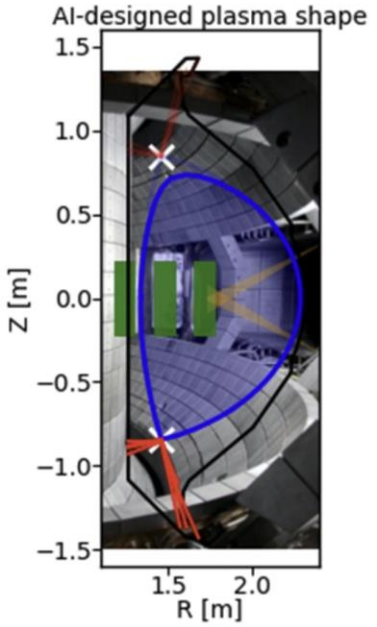
$$\mathcal{J} := \int_{\Omega \times [0,T]} (|\beta_p(t,x) - \beta_p^*(x)|^2 + |l_i(t,x) - l_i^*(x)|^2) \mathrm{d}x\mathrm{d}t$$

Safety score:

$$s := - \inf_{(t,x) \in [0,T] \times \Omega} \{q_{95}(t,x)\}$$

Methods	$\mathcal{J} \downarrow$	$\mathcal{R}_{\text{sample}} \downarrow$	$\mathcal{R}_{\text{time}} \downarrow$
BC	0.0610	42%	1.34%
BC-Safe	0.0811	4%	0.03%
SL-Lag	0.8812	0%	0.00%
MPC-Lag	0.8659	0%	0.00%
CDT	0.0071	8%	0.54%
TREBI	0.0261	0%	0.00%
SafeDiffCon	**0.0121**	0%	0.00%





Thank you!

If you have any questions, please feel free to contact us at:

hupeiyan18@mails.ucas.ac.cn

wutailin@westlake.edu.cn

Group Website: