Conformal Reliability:
A New Evaluation Metric for Conditional Generation


Yachen Gao


Xinwei Sun


Yikai Wang


Ye Shi


Jingya Wang


Jianfeng Feng

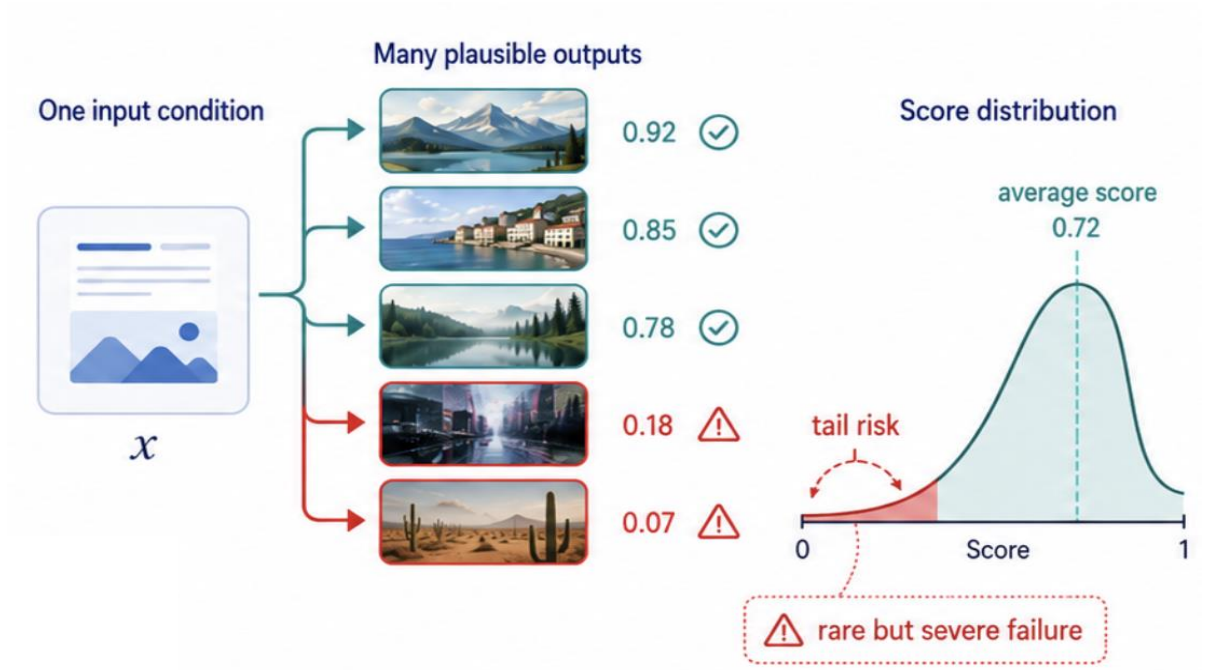

Yanwei Fu

Fudan University · Shanghai Innovation Institute ·
Nanyang Technological University · ShanghaiTech University

Why Average Score Are Not Enough

Single-output evaluation misses uncertainty.

- Conditional generation is stochastic
- A high average score can hide rare failures
- Reliability should measure the performance floor
- This matters in safety-critical applications



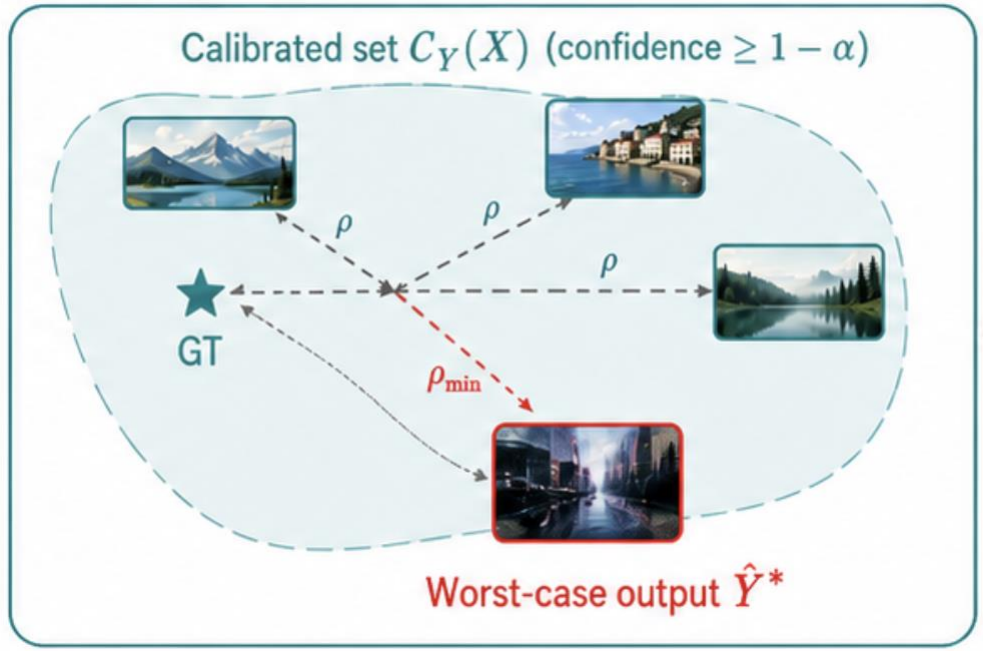
⚠ Average quality does not reveal how bad a plausible output can be.

Reliability Score

Turn any similarity metric into an uncertainty-aware worst-case metric.

- **Step 1: Build a prediction set**
  - Construct $C_Y(x)$ at confidence level $1-\alpha$.
  - This set contains statistically plausible outputs.
- **Step 2: Evaluate the worst metric value**
  - Compute the minimum metric value inside the set.
  - This yields a lower bound on performance.

$$\min_{\hat{Y} \in C_Y(X)} \rho(\hat{Y}, \mathrm{GT}) \quad \text{subject to} \quad \mathbb{P}(\hat{Y} \in C_Y(X)) \geq 1-\alpha$$



$C_Y(X)$: calibrated prediction set

ρ: task-specific similarity metric

$1-\alpha$: confidence level

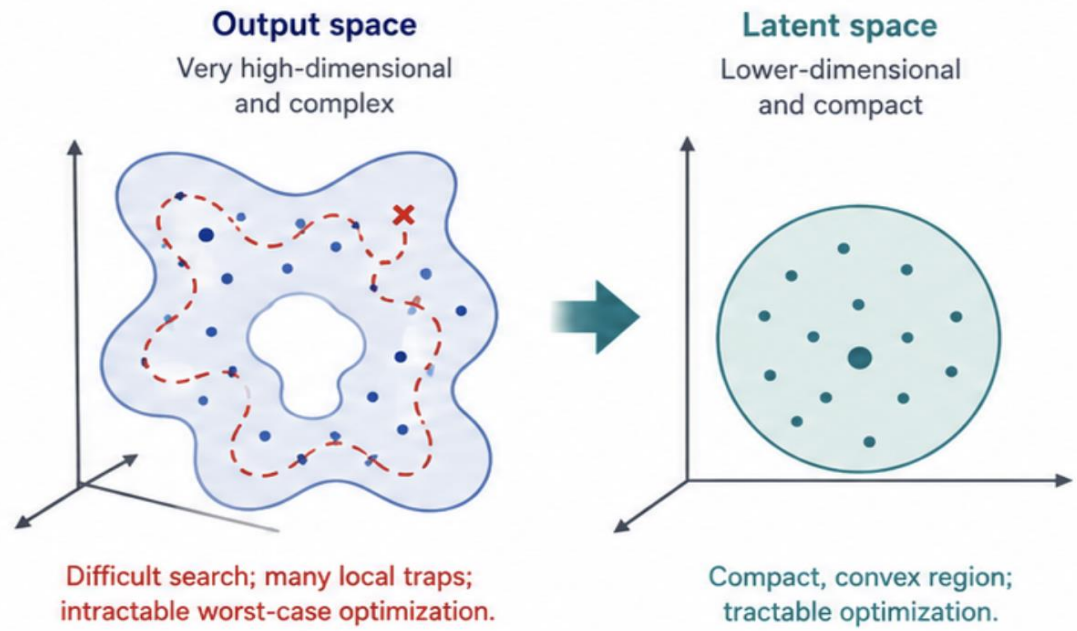
Output: worst-case performance floor

Intuition: Among statistically plausible outputs, how low can performance go?

Why Is This Hard

Direct reliability optimization is intractable in output space.

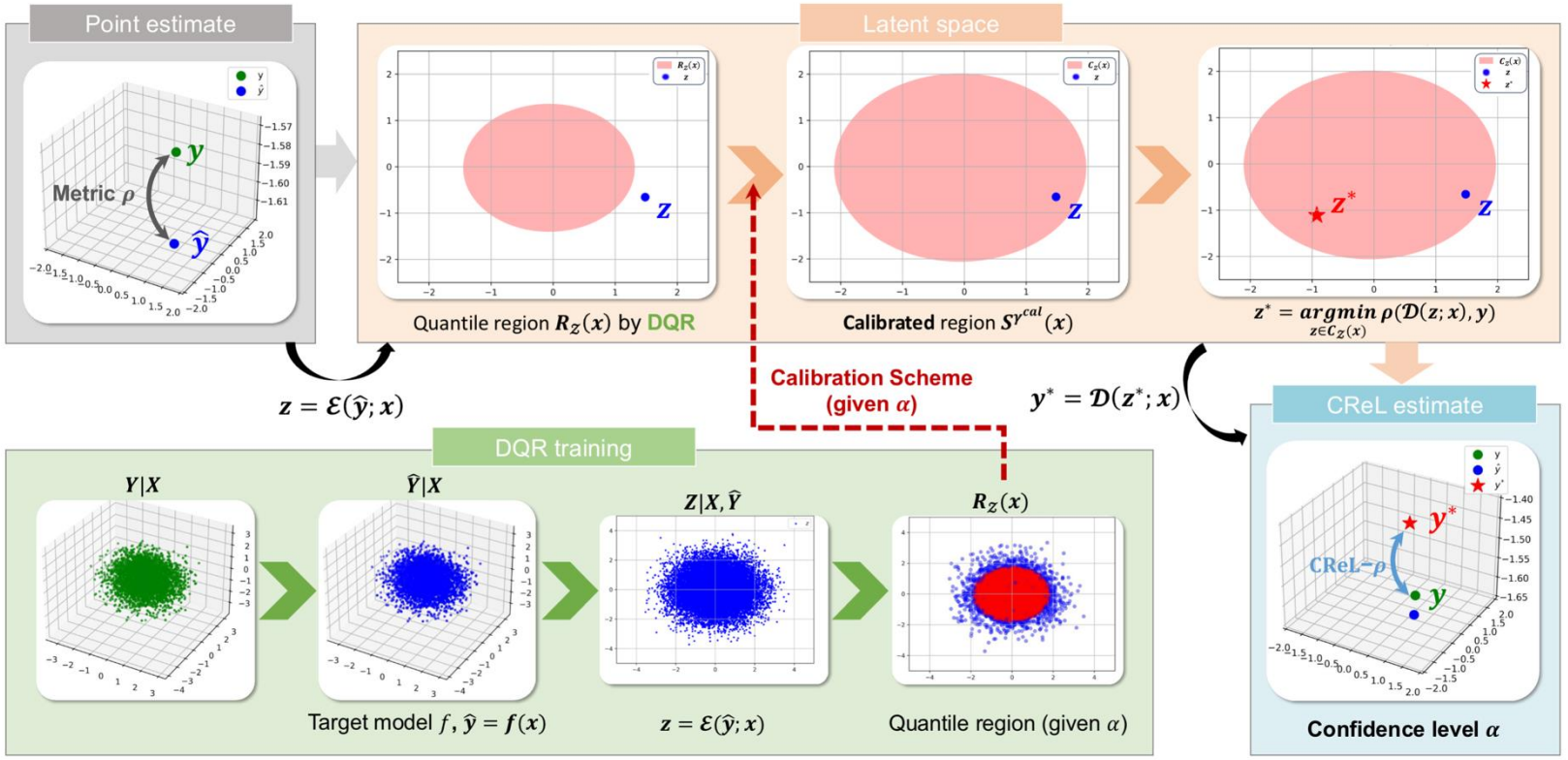
- **High-dimensional outputs**
  - Images and texts live in complex spaces.
- **Nonconvex prediction sets**
  - Plausible outputs do not form simple regions.
- **Nonconvex metrics**
  - Similarity functions can be difficult to optimize.
- **Direct worst-case search**
  - Brute-force optimizations computationally expensive.



🎯 We need a representation where calibration and optimization become tractable.

CReL: Conformal ReLiability

Calibrate and optimize in latent space, then decode back to the output space.



More informative prediction sets

Regions shaped by the data in latent space.

Coverage-aware evaluation

Conformal guarantees at level $1 - \alpha$.

Efficient optimization in a convex latent region

Convexity enablee scalable worst-case search.

✅CReL computes reliability by moving calibration and optimization into a structured latent space.

Guarantees and Computation

CReL offers statistical validity and a tractable optimization view.

Latent-space coverage

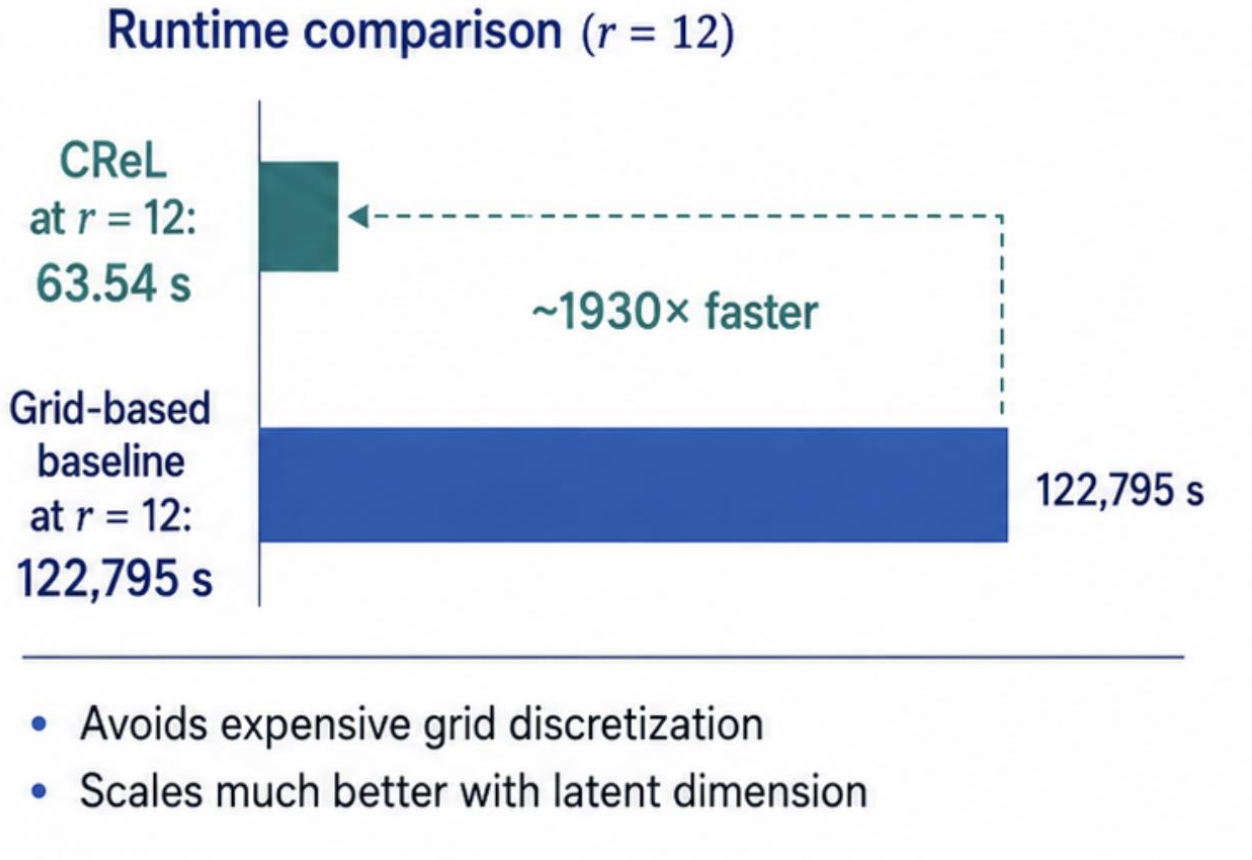
After calibration, the latent prediction set attains the target coverage level.

Decoded-set validity

If the latent generative model recovers the conditional output distribution, the decoded set also satisfies coverage.

Convex feasible region

The calibrated latent set is convex and compact, enabling projected optimization.



- Avoids expensive grid discretization
- Scales much better with latent dimension

Results

Reliability changes how we compare generative models.

Table 1. Coverage ratio and area on the *nonlinear* synthetic dataset with different nominal levels α.

α	Coverage					Area in $\mathcal{Y}$		
	Ours-$\mathcal{Z}$	Ours-$\mathcal{Y}$	Feldman-$\mathcal{Y}$	DQR-$\mathcal{Z}$	DQR-$\mathcal{Y}$	Ours	Feldman	DQR
0.02	0.9770	0.9760	0.9718	0.9818	0.9872	398.5	377.8	749.1
0.10	0.8953	0.8915	0.8940	0.8823	0.9145	232.7	234.5	287.4

- More informative prediction sets.
- CReL effectively identifies misalignments.

Table 2. Quantitative results of the image-to-text generation task at $\alpha = 0.1$, with differences between CReL-ρ and ρ (Δ) highlighted in blue. Superscripts indicate the performance rank.

Model	CLIP-SIM		BERT-SIM	
	CLIP	CReL-CLIP	BERT	CReL-BERT
BLIP-base	0.2330^{4}	0.0070^{1} (-0.2260)	0.8349^{3}	0.6335^{3} (-0.2014)
BLIP-large	0.2453^{3}	-0.0074^{4} (-0.2527)	0.8106^{4}	0.5631^{4} (-0.2475)
GIT-base	0.2511^{2}	-0.0021^{2} (-0.2532)	0.8620^{2}	0.6474^{1} (-0.2146)
GIT-large	0.2550^{1}	-0.0043^{3} (-0.2593)	0.8649^{1}	0.6459^{2} (-0.2190)

	CLIP	CReL-CLIP
GT caption: A baby in a bouncy seat chewing on a plastic toy.		
BLIP-base: a baby in a car seat	0.1920^{4}	-0.0202^{2}
BLIP-large: there is a baby sitting in a high chair with a toy in his mouth	0.2306^{2}	-0.0039^{1}
GIT-base: my son in his high chair	0.2529^{1}	-0.0443^{4}
GIT-large: sitting in a chair with a red toy	0.2213^{3}	-0.0435^{3}

	BERT	CReL-BERT
GT caption: Three cell phones lying next to each other on a wooden table.		
BLIP-base: a group of cell phones sitting on a table	0.8825^{3}	0.6388^{4}
BLIP-large: three cell phones are sitting on a table with a wooden surface	0.7509^{4}	0.6560^{2}
GIT-base: three cell phones sitting on top of a wooden table.	0.9880^{1}	0.6627^{1}
GIT-large: three cell phones sitting on a table.	0.9788^{2}	0.6421^{3}

Figure 4. Qualitative results of image-to-text models ($\alpha = 0.1$). Superscripts denote rank.

Takeaways

Reliability changes how we compare generative models.

- **More informative sets**
  - CReL achieves tighter calibrated regions than standard alternatives.
- **Ranking changes matter**
  - Worst-case evaluation can reverse conclusions drawn from average scores.
- **Risk-aware evaluation**
  - CReL quantifies worst-case performance under calibrated uncertainty.

THANKS!

Use CReL when worst-case behavior matters.