PITT | SWANSON ENGINEERING
INDUSTRIAL

HARVARD
School of Engineering and Applied Sciences

CAMCA
Center for Advanced Medical Computing and Analysis

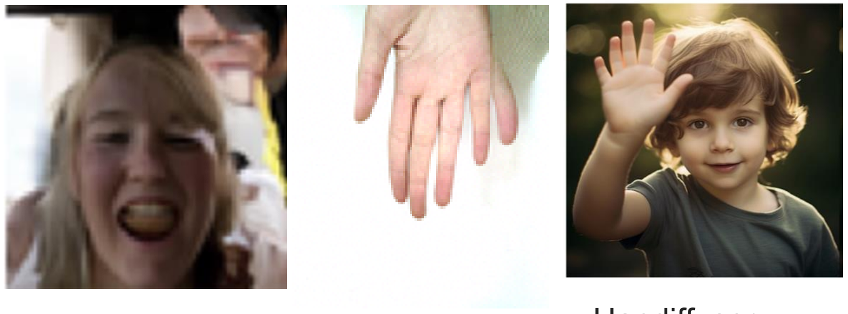
RODS: Robust Optimization Inspired Diffusion Sampling for Detecting and Reducing Hallucination in Generative Models

Yiqi Tian [*1,2], Pengfei Jin[*1], Mingze Yuan[1,3], Na Li[3], Bo Zeng [†2], and Quanzheng Li[†1]

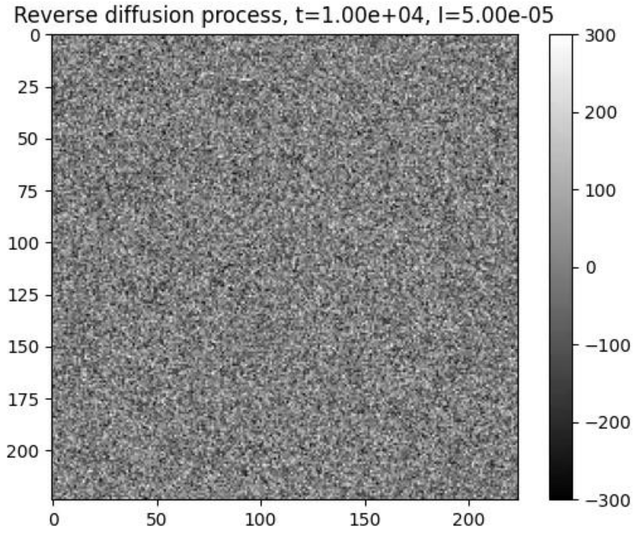
[1]Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114
[2]Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15261
[3]School of Engineering and Applied Sciences, Harvard University, Boston, MA 02138

Background and Motivation

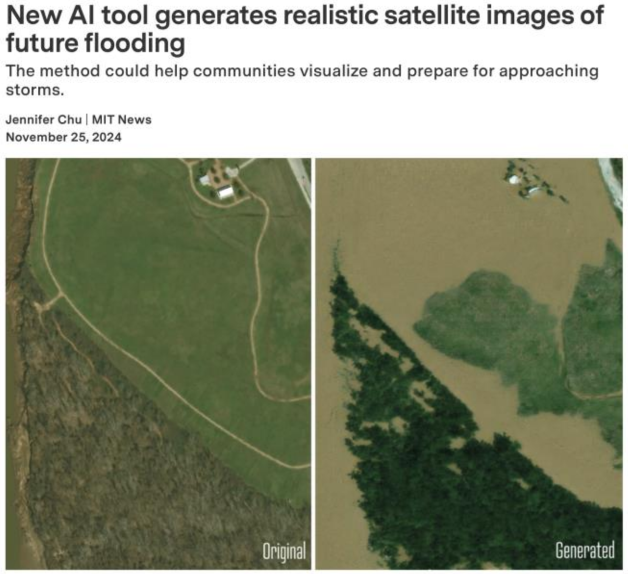
- **Hallucination in Diffusion Model**



Handiffuser,
Narasimhaswamy, S et al.,
CVPR 2024



Reverse diffusion process, t=1.00e+04, l=5.00e-05



New AI tool generates realistic satellite images of future flooding

The method could help communities visualize and prepare for approaching storms.

Jennifer Chu | MIT News
November 25, 2024

Related Work

Fine-tuning

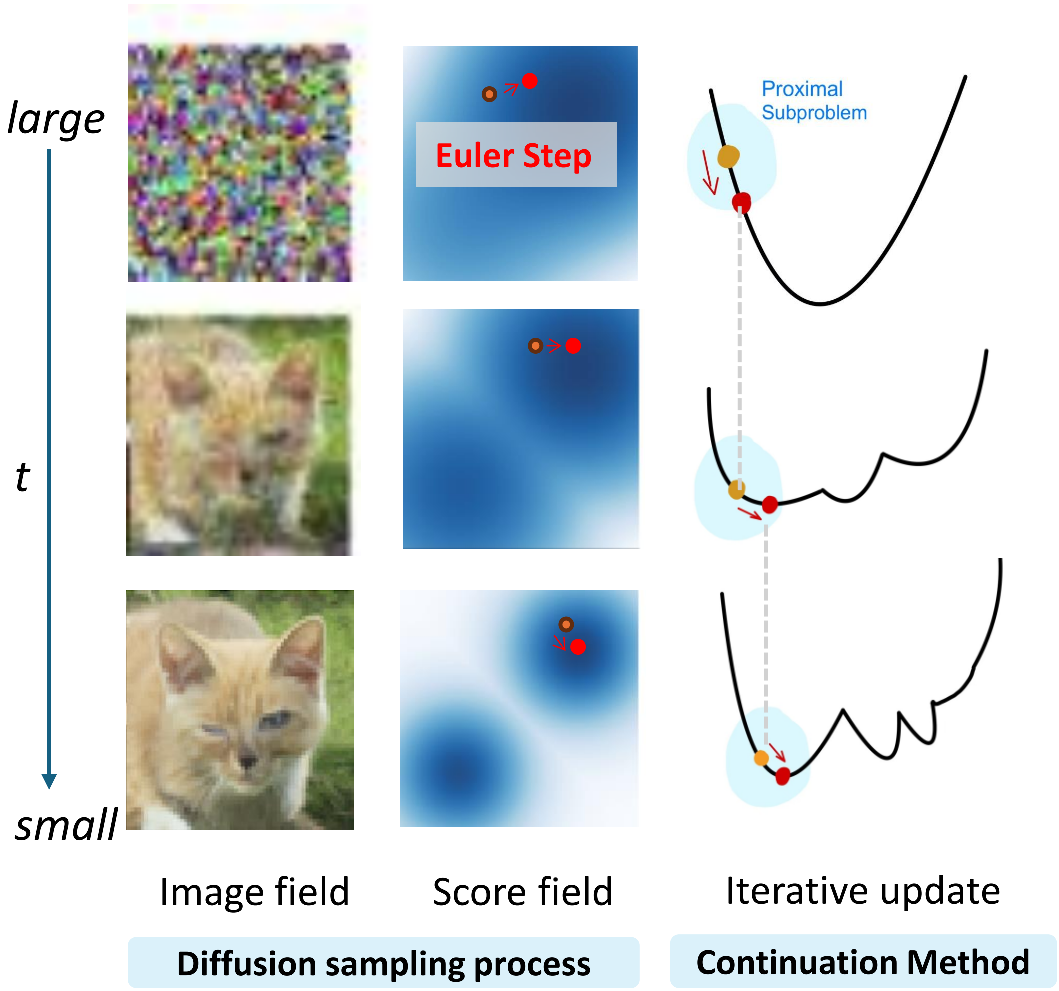
Filtering

heuristic modification

- **Limitation**
  - **Lack of theoretical guidance**
  - **Lack of benchmark**

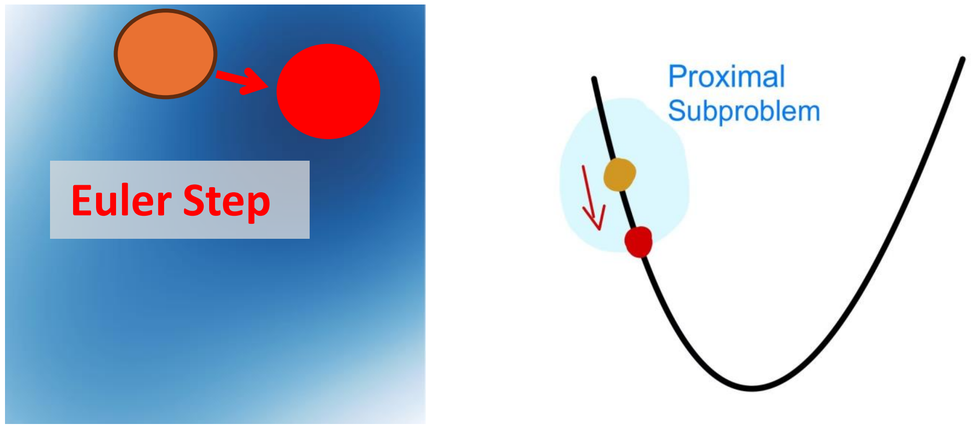
Theory & Methods

An Optimization View of Diffusion Sampling Process



Diffusion sampling via ODE = continuation method

Euler = One-Step Gradient = Proximal Update

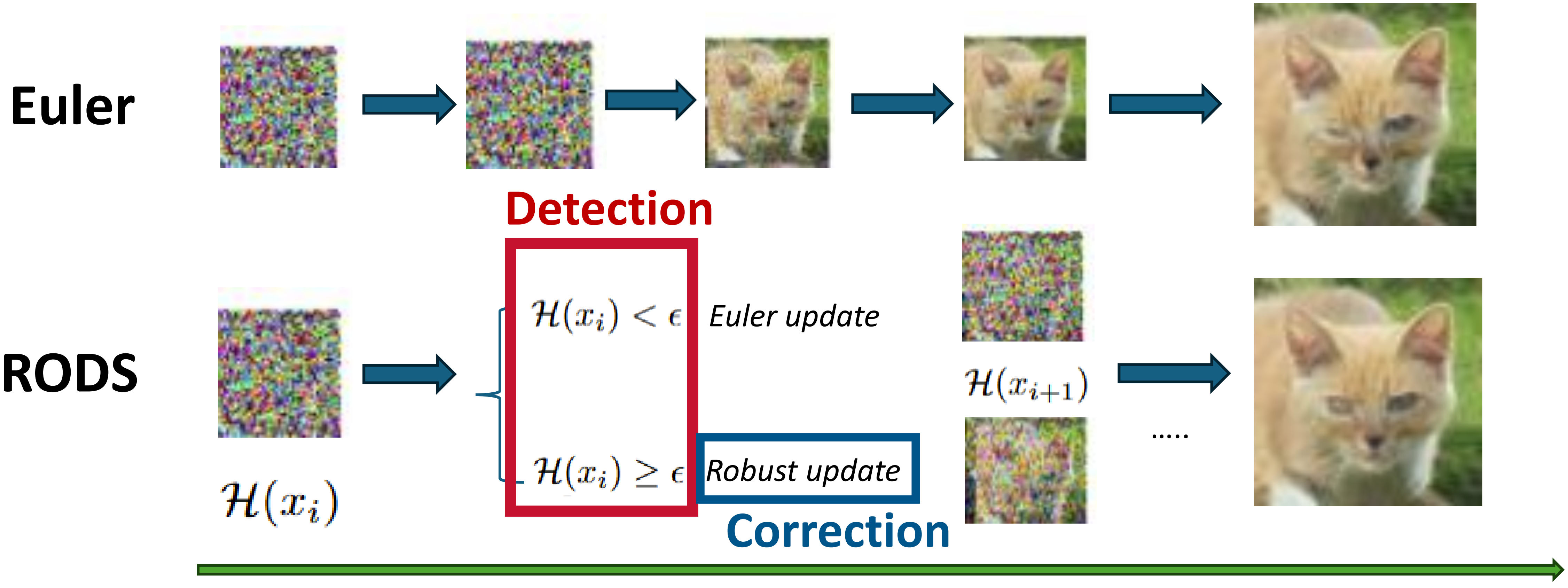


Inspiration: Leveraging various optimization tools

RODS: Robust Optimization inspired Diffusion Sampler

Uncertainty

Hallucination



Euler

Detection

RODS

$\mathcal{H}(x_i) < \epsilon$ *Euler update*

$\mathcal{H}(x_i) \geq \epsilon$ *Robust update*

$\mathcal{H}(x_i)$

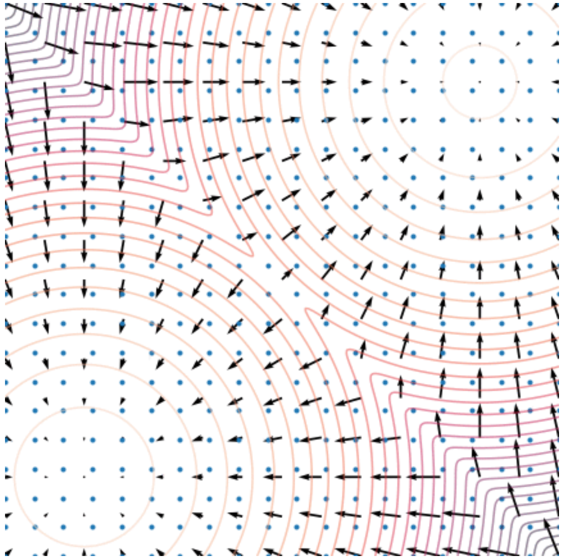
$\mathcal{H}(x_{i+1})$

.....

Correction

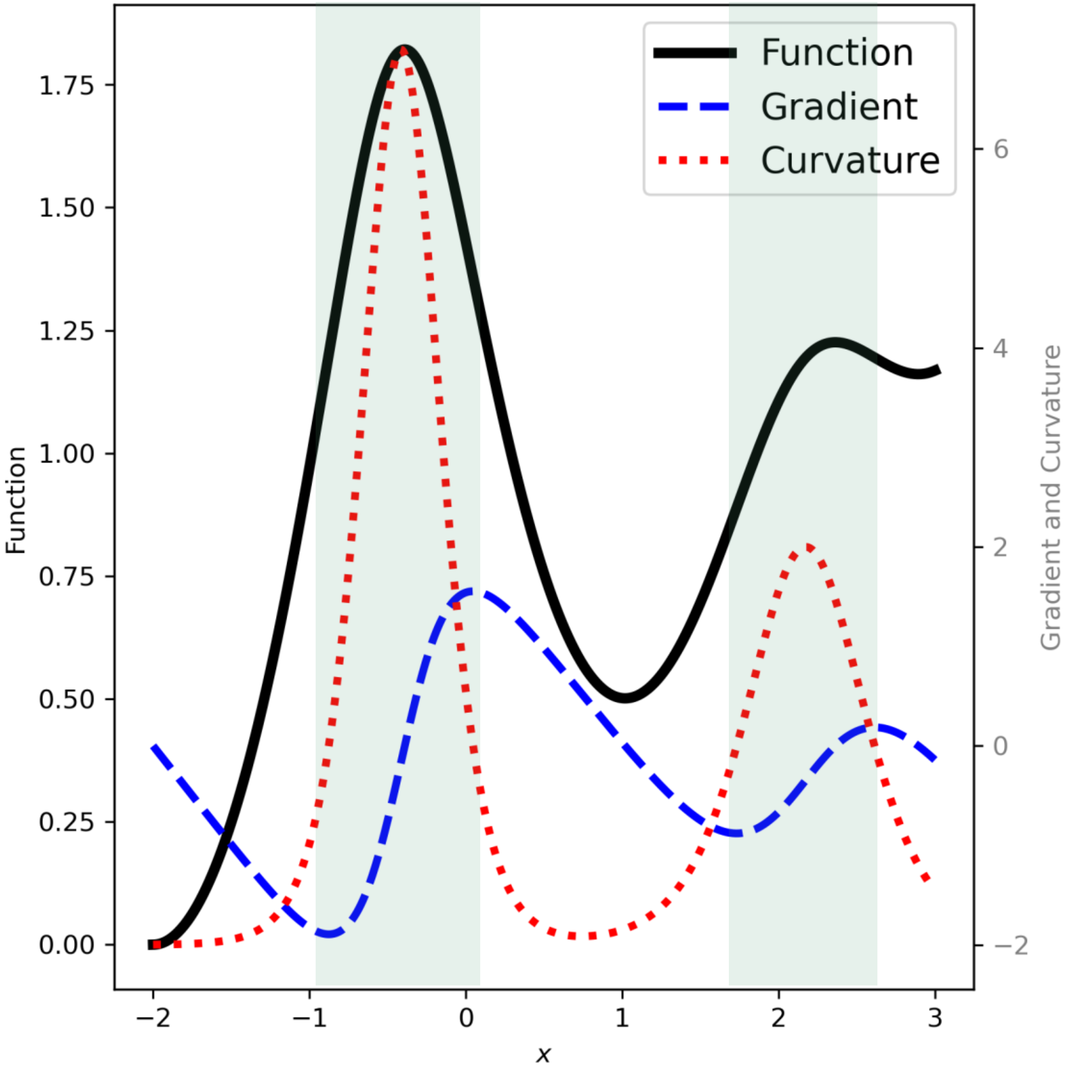
Time step i

RODS: Robust Optimization inspired Diffusion Sampler



Low-density region
Unstable score field

Rapid gradient change
High curvature



- **Curvature Change Detection**
  - Hallucination Index:

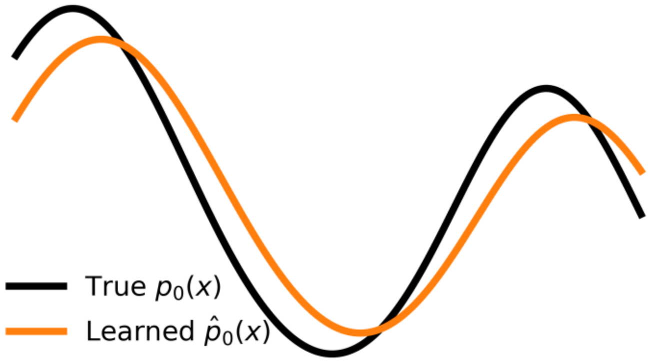
$$\mathcal{H}(x) = \|\nabla_x \|v(x+\delta)\| - \nabla_x \|v(x)\|\|,$$

$$\delta = \arg \max_{\|\delta\|=\rho} \|v(x+\delta)\|,$$

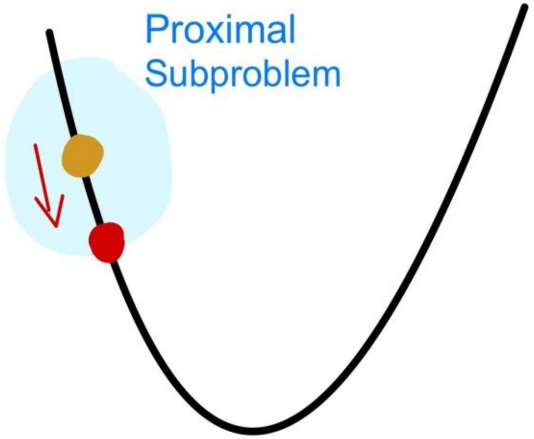
RODS: Robust Optimization inspired Diffusion Sampler

- **Robust Sampling Process**

$$\min_{x} f_t(x) \iff \min_{x}\left[-\log p_t(x)\right].$$



$$\min_{x} \max_{\hat{f}_t \in \mathcal{F}_t} \hat{f}_t(x) \iff \min_{x} \max_{\hat{p}_t \in \mathcal{P}_t}\left[-\log \hat{p}_t(x)\right]$$

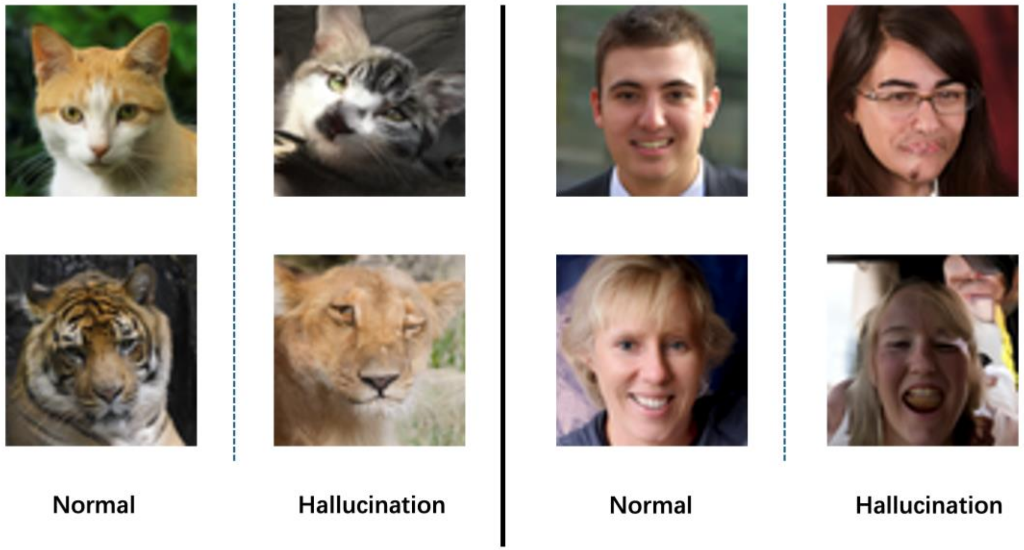


Euler =
One-Step Gradient =
Proximal Update

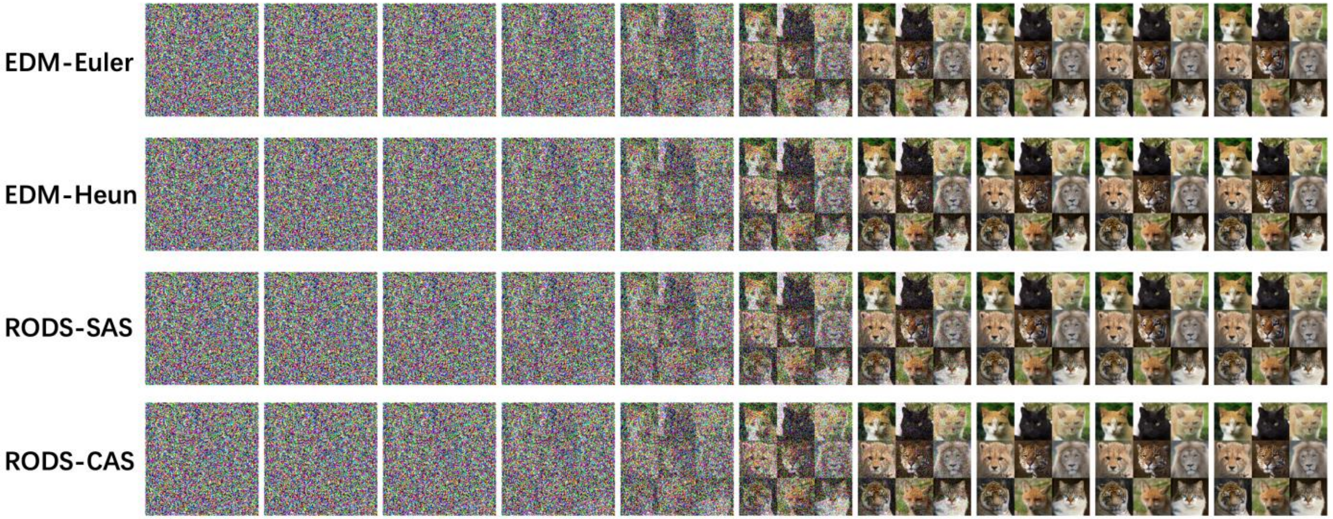
Experiments & Conclusion

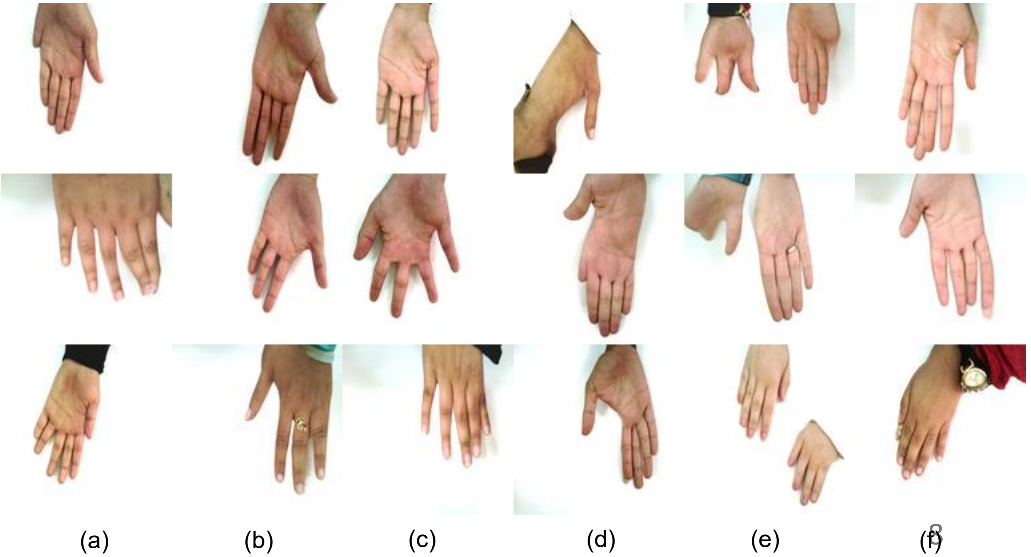
Experiment Setting

- Datasets:
  - AFHQ-v2
  - FFHQ
  - 11k-hands



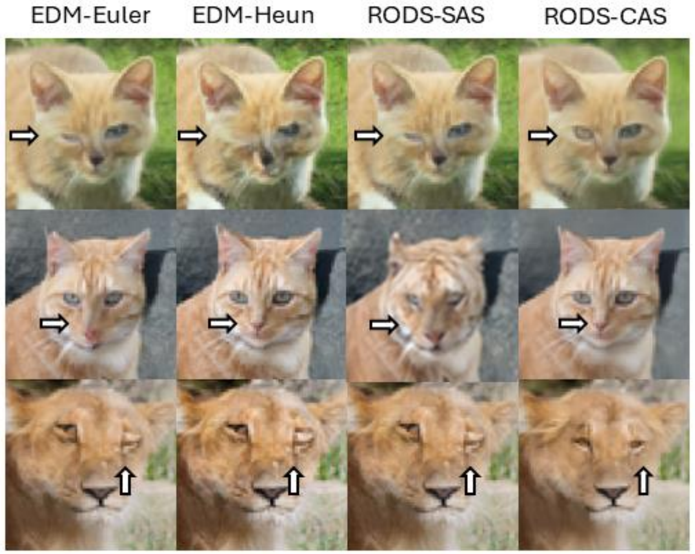
- Labeling by human

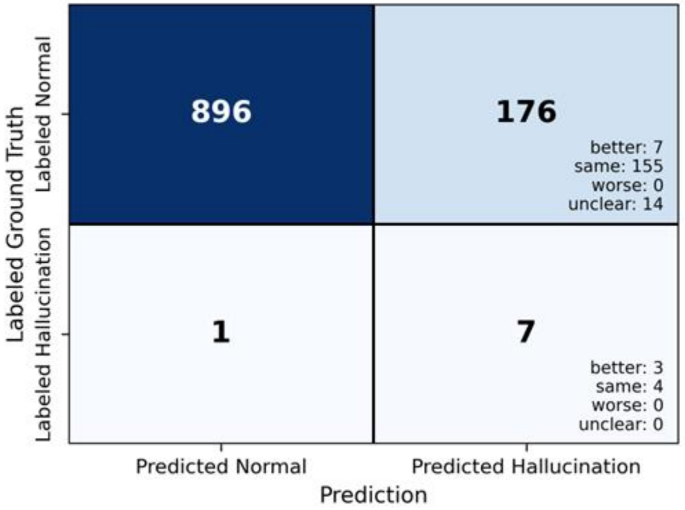




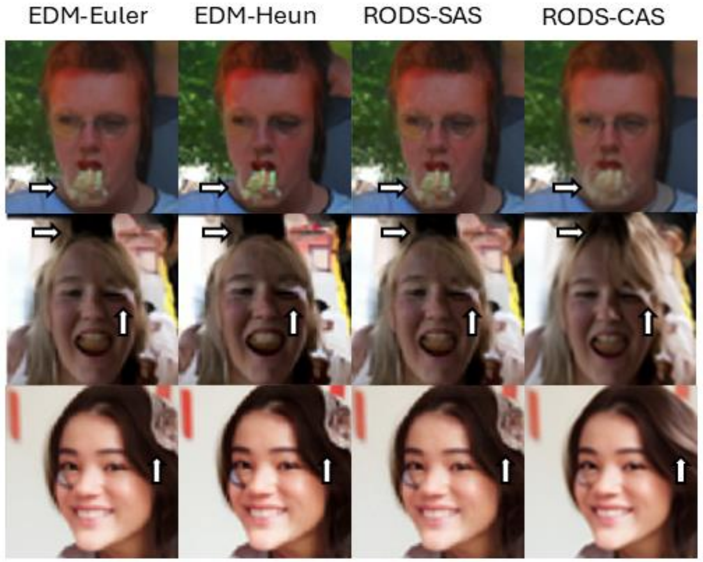
Experiment Results

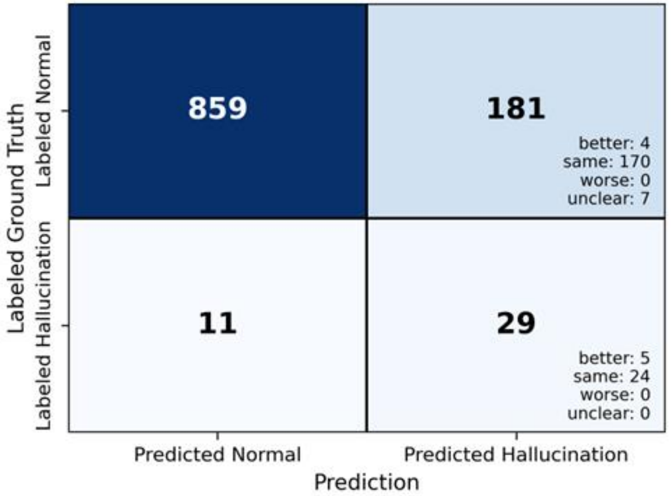
- **Effective Detection**
  - Animal Face: 87.5%
  - Human Face: 72.5%
  - Hands: 96.6%
- **Reduces hallucination**
  - Animal Face: 43%
  - Human Face: 25%
  - Hands: 27%
- **No new hallucinations introduced**





	Predicted Normal	Predicted Hallucination
Labeled Normal	896	176 better: 7 same: 155 worse: 0 unclear: 14
Labeled Hallucination	1	7 better: 3 same: 4 worse: 0 unclear: 0





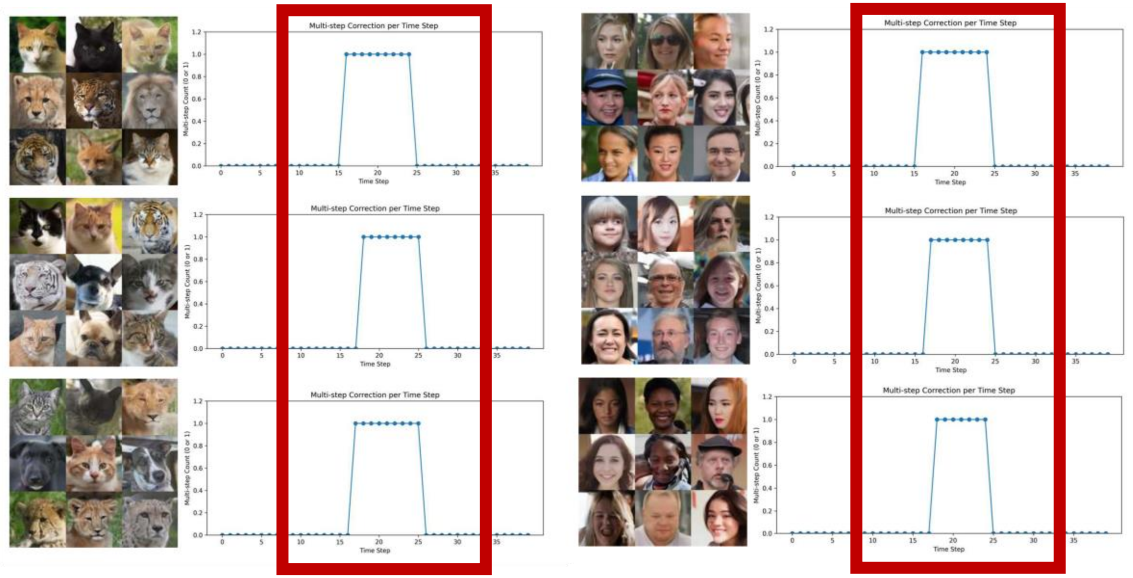
	Predicted Normal	Predicted Hallucination
Labeled Normal	859	181 better: 4 same: 170 worse: 0 unclear: 7
Labeled Hallucination	11	29 better: 5 same: 24 worse: 0 unclear: 0

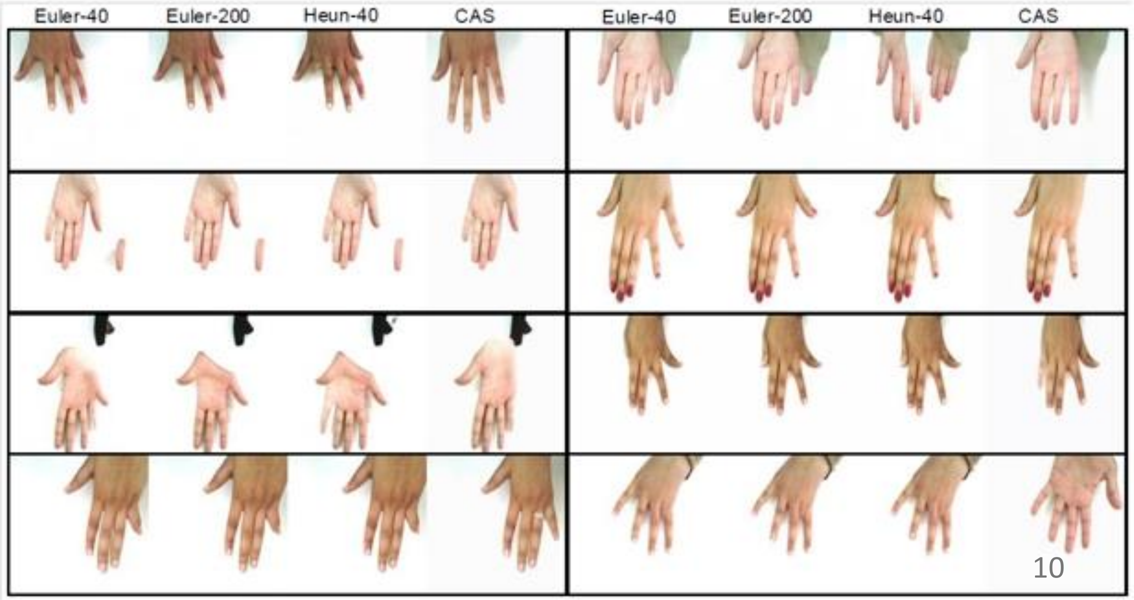
Experiment Results

- **Hallucination occur in the middle**

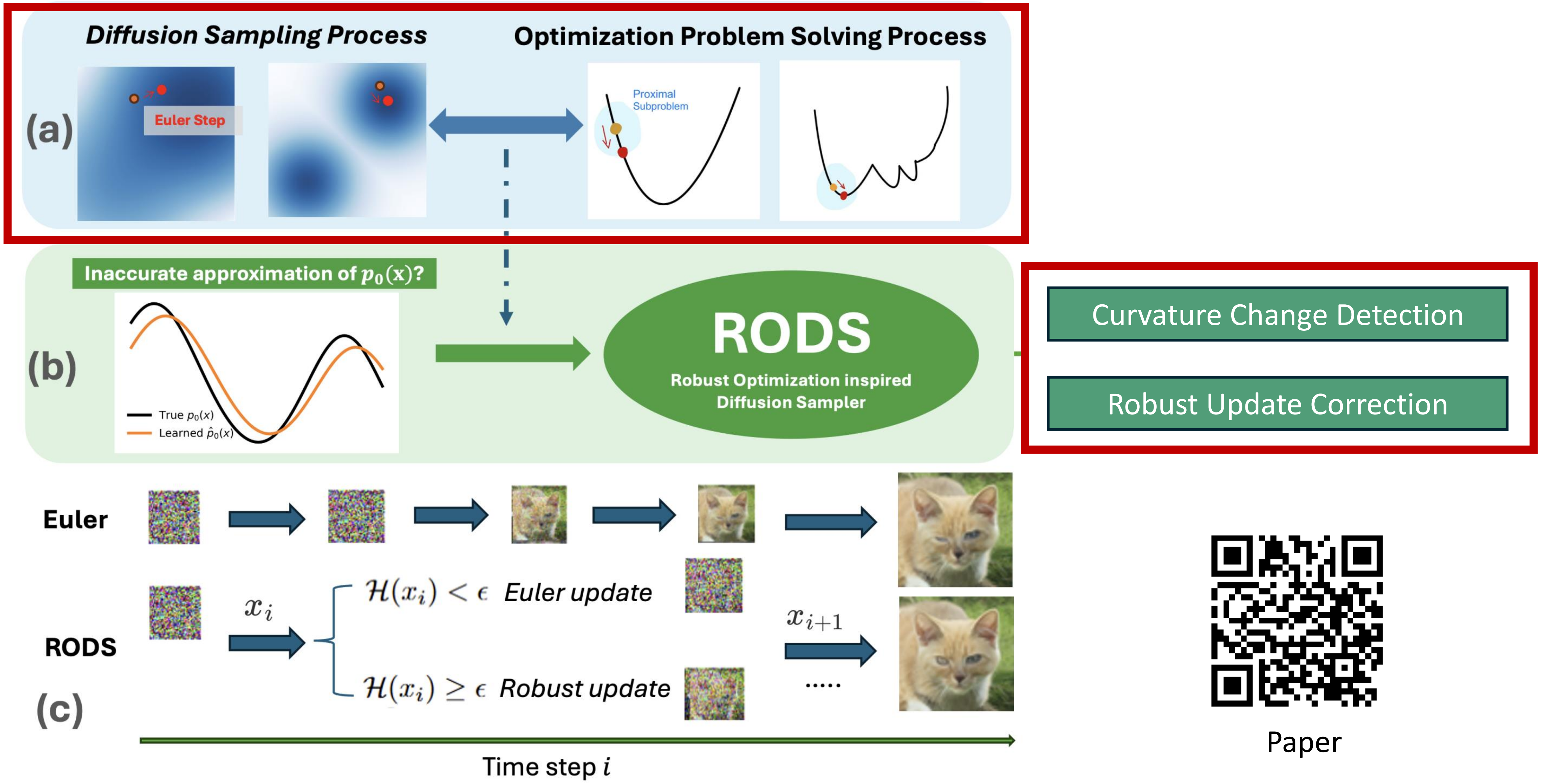
- **Keep Comparable Time**

Metric	EDM-Euler	EDM-Heun	RODS-CAS
Hall.% ⬇	19.4%	18.0%	**14.3%**
Correction % ⬆	N/A	17.7%	**26.3%**
New Hall. % ⬇	N/A	2.5%	**0.0%**
Time (s)	**2.71**	5.38	4.82





Conclusion


(a)
Diffusion Sampling Process
Euler Step
Optimization Problem Solving Process
Proximal Subproblem
(b)
Inaccurate approximation of $p_0(\mathbf{x})$?
True $p_0(x)$
Learned $\hat{p}_0(x)$
RODS
Robust Optimization inspired Diffusion Sampler
Curvature Change Detection
Robust Update Correction
(c)
Euler
RODS
x_i
$\mathcal{H}(x_i) < \epsilon$ Euler update
$\mathcal{H}(x_i) \geq \epsilon$ Robust update
x_{i+1}
.....
Time step i

Paper