Tackling **Structural Hallucination** in Image Translation with Local Diffusion

Seunghoi Kim*, Chen Jin*, Tom Diethe, Matteo Figini, Henry F. J. Tregidgo, Asher Mullokandov, Philip Teare, and Daniel C. Alexander

Introduction

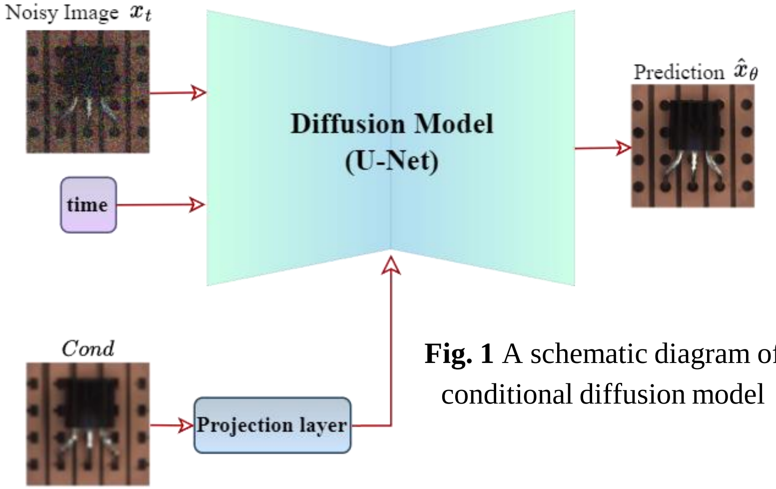


Fig. 1 A schematic diagram of conditional diffusion model

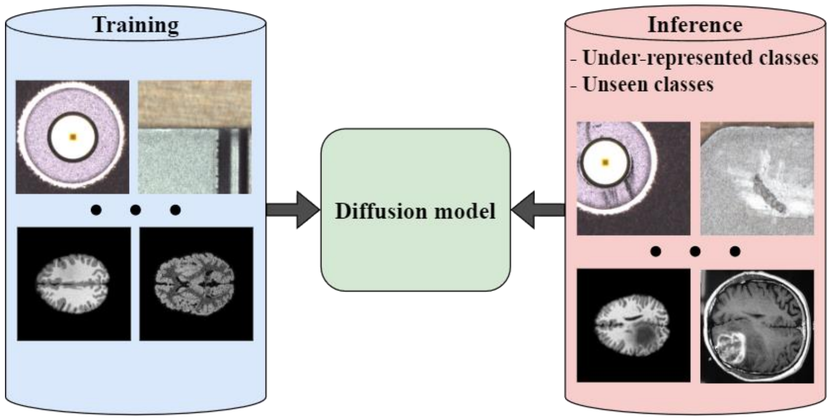


Fig. 2 Illustration of data-shift problem

- Conditional diffusion models have attained state-of-the-art performances in various image translation tasks
- Generative model f_θ learns joint distribution of train data and its condition
- Many applications with significant under-represented classes (e.g. rare diseases, defects) exist
- Performs well on in-distribution (IND) data, but what if the condition contains out-of-distribution (OOD) region?

$$Error(f_\theta(c_{ood\&ind})) \gg Error(f_\theta(c_{ind})) \text{ or at worst hallucination!}$$

Problem: Hallucination

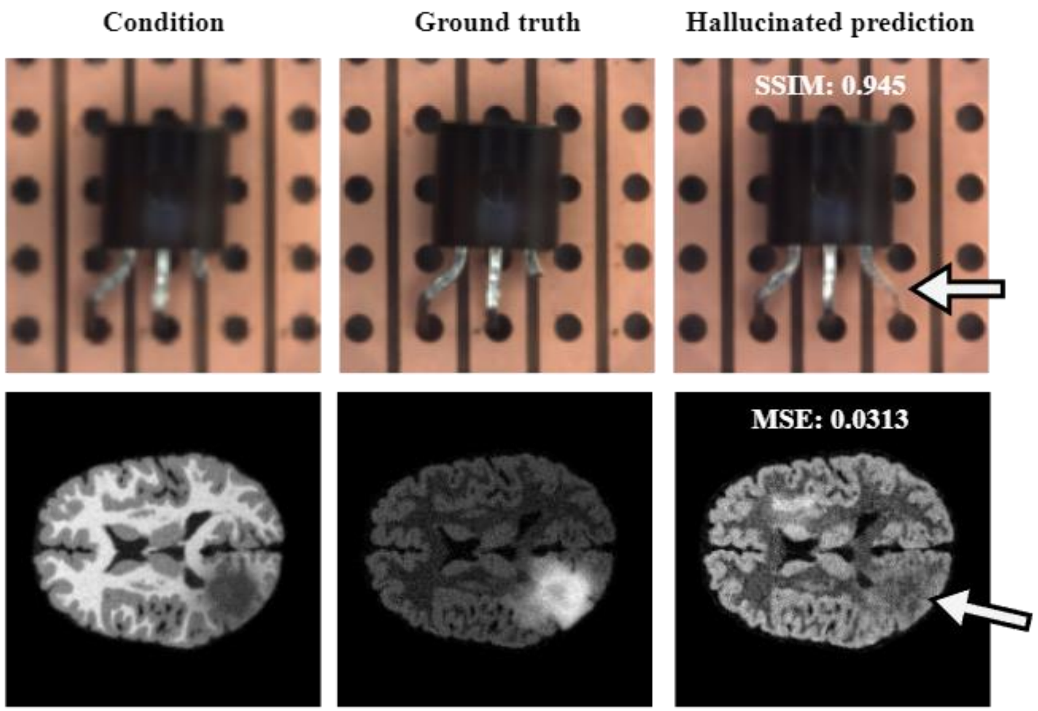


Fig. 3 Visual examples of structural hallucination

- What is structural hallucination and why should we care?
  - Realistic-looking but inaccurately reconstructed features, leading to discrepancies with the actual structure
  - Misinterpretation => patient misdiagnosis, machine failure, increase in time and cost
  - Often **insensitive** using **standard image quality metrics** (e.g. MSE, SSIM)

Solution?

- Simple way to fix => fine-tuning, but **expensive**

Hypothesis & Verification

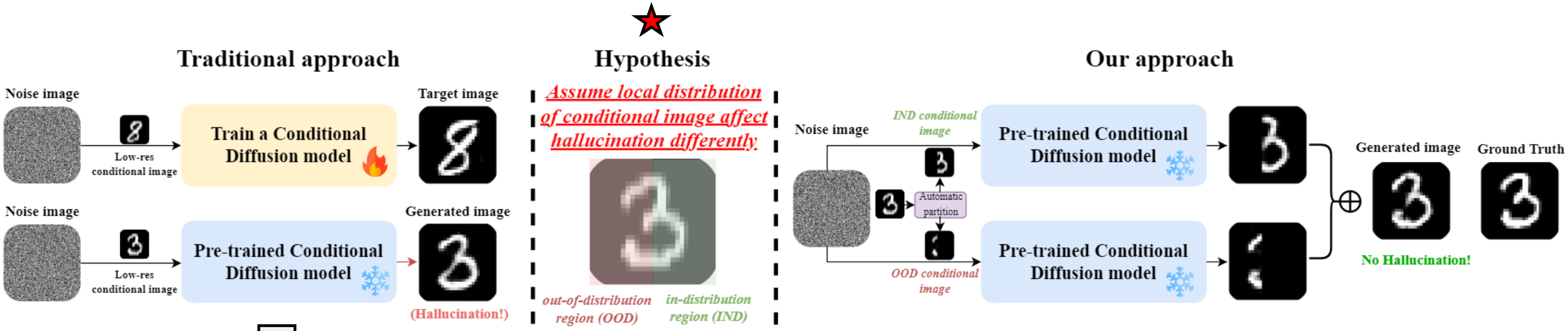
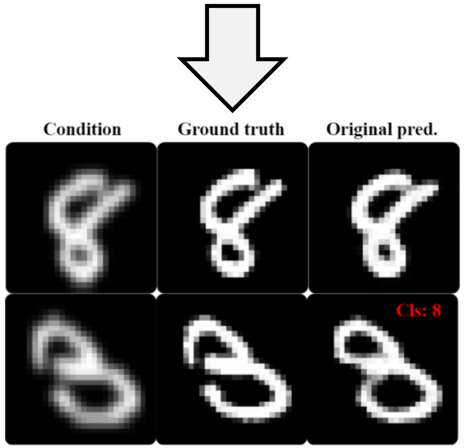


Fig. 4 Traditional conditional diffusion process vs. our OOD/IND Local Diffusion



<u>**Traditional approach leads to hallucination!**</u>

Can OOD-based Local Image Generation Help to Reduce Hallucination?

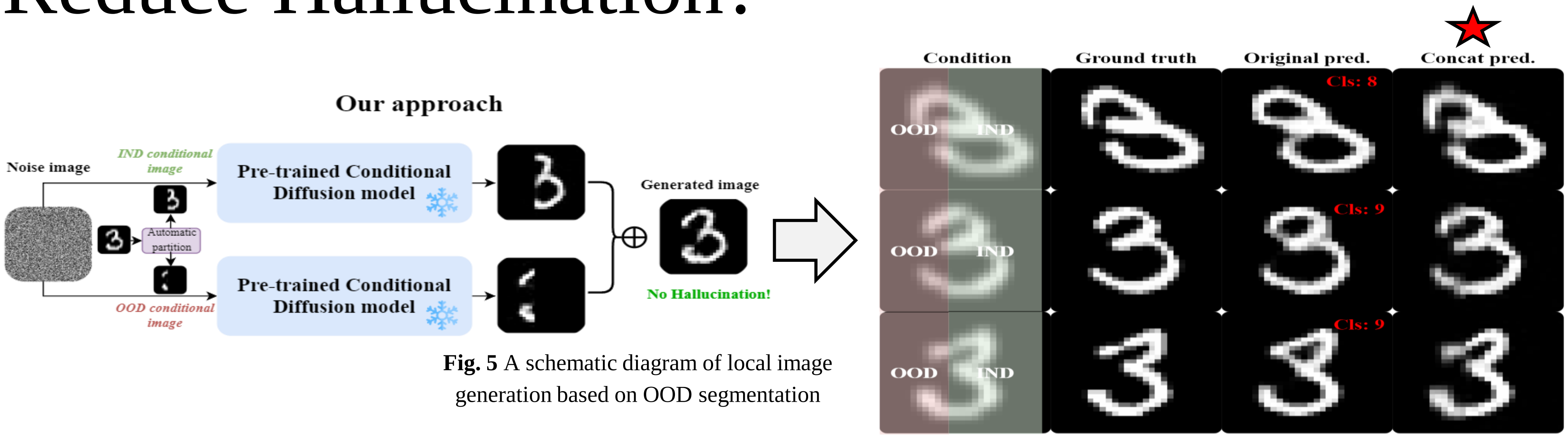
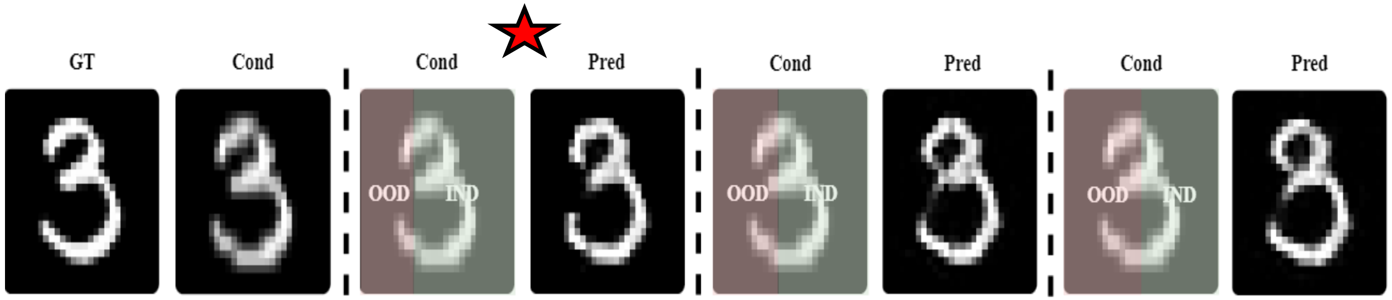


Fig. 5 A schematic diagram of local image generation based on OOD segmentation



Precise segmentation of OOD is crucial for hallucination mitigation!

Fig. 6 Visual illustration of the impact of shifting OOD boundary

Identifying Hallucination Hotspots in Diffusion Models

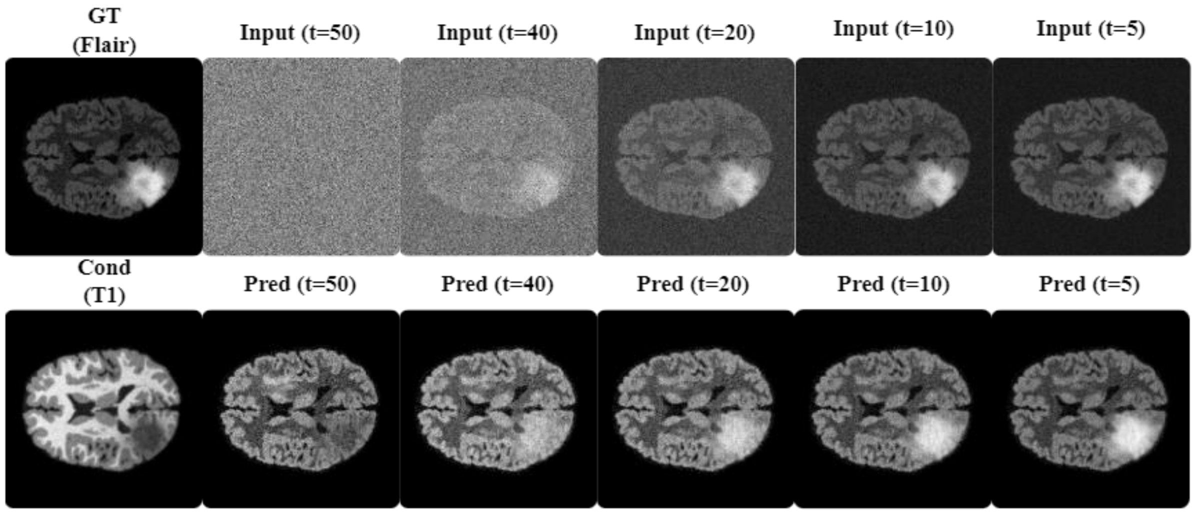


Fig. 7 Qualitative comparisons of predictions starting from different intermediate time points. We sample noisy GT (Flair) and perform a reverse process from it by conditioning the corresponding T1 image.

Less hallucination!

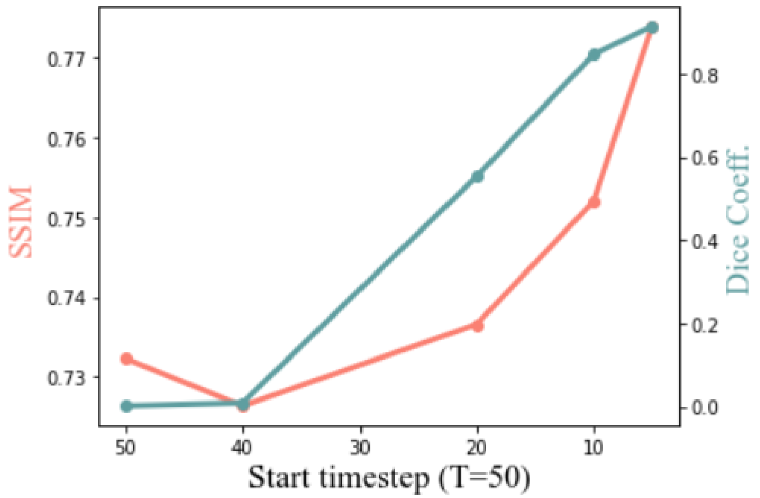


Fig. 8 Dice score of tumor segmentation and SSIM of predicted images

Methods: Local Diffusion

-	Dice	Train/Inf. time (s)	Gradient descent
CFA[4]	0.698	140.8/**0.017**	Yes
FastFlow[3]	0.670	166.1/0.032	Yes
PatchCore	**0.789**	**13.2**/**0.017**	No

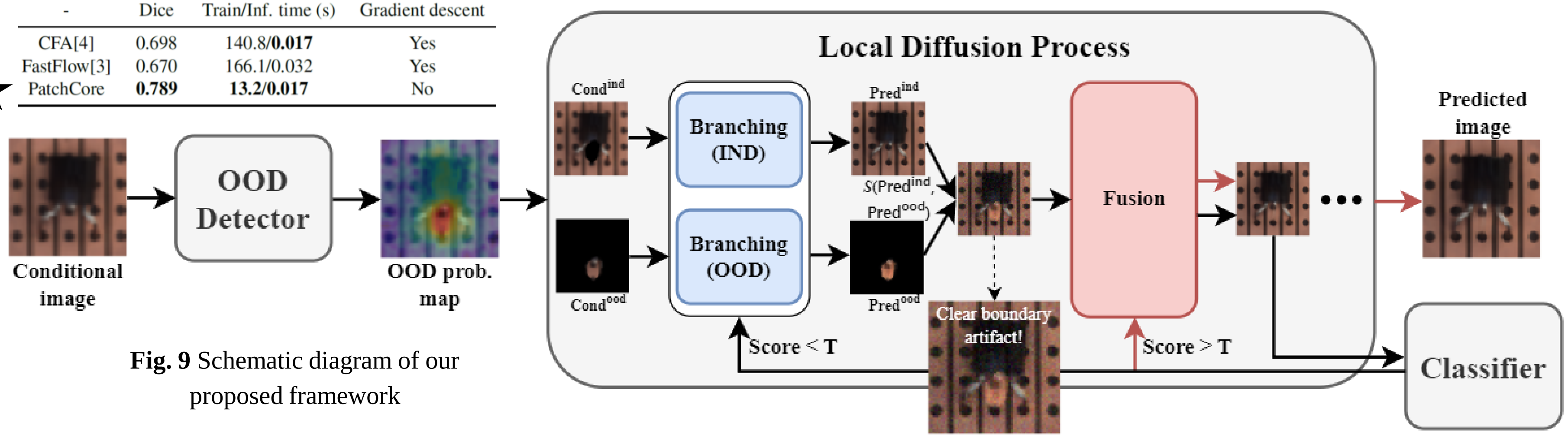


Fig. 9 Schematic diagram of our proposed framework

- ➢ **OOD estimation**: One-class classification anomaly detector (PatchCore [CVPR’22])
- ➢ **Branching**: Separate local image generation based on OOD probability map
- ➢ **Fusion**: Fuse the OOD/IND predictions for more cohesive image generation
- ➢ **Classifier**: Checks if the prediction at intermediate time step contains hallucination

Main Results (Quantitative)

	MNIST		BraTS		MVTec AD	
	PSNR (↑)	*SSIM* (↑)	*PSNR* (↑)	*SSIM* (↑)	*PSNR* (↑)	*SSIM* (↑)
DDPM [11]	20.8±1.90	0.897±0.04	19.5±1.75	0.709±0.04	26.8±3.10	0.839±0.10
DDIM [32] (0.1T)	20.8±1.88	0.895±0.04	19.8±1.53	0.715±0.05	27.3±3.00	0.844±0.10
DDIM [32] (0.5T)	20.6±1.86	0.899±0.04	19.7±1.61	0.703±0.04	27.1±2.98	0.840±0.10
DDPM + DSI [38]	20.8±1.90	0.898±0.04	18.7±1.74	0.695±0.04	26.6±3.40	0.815±0.10
DDIM + Ours (0.1T)	**20.9±1.87**	0.897±0.04	20.7±1.52	**0.720±0.05**	**27.5±3.02**	**0.847±0.09**
DDIM + Ours (0.5T)	20.8±1.82	**0.902±0.03**	20.9±1.61	0.711±0.04	27.4±2.97	0.844±0.10
DDPM + Ours	**20.9±1.85**	0.900±0.04	**21.2±1.74**	**0.720±0.03**	27.0±3.05	0.843±0.10
p-value	0.014	0.016	≈ 0	0.02	0.001	0.004

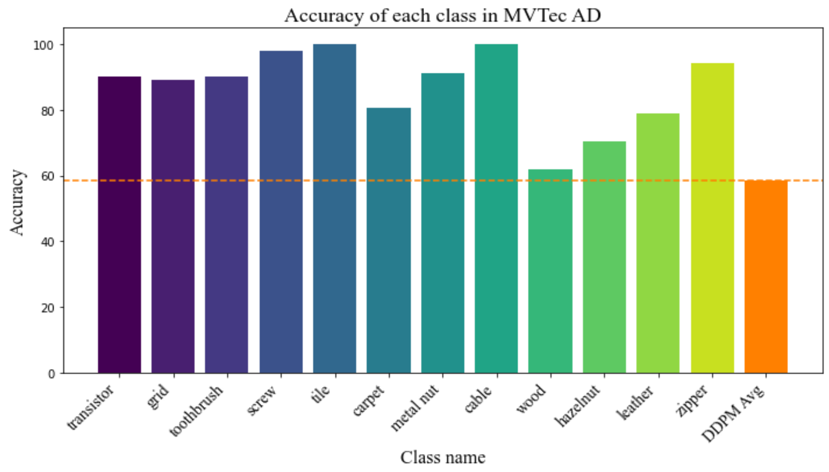
Tab. 1 Quantitative comparisons of overall image quality across various datasets, where an upward arrow signifies that a higher value is better. T represents the total number of time steps for sampling.

Downstream task performance is more important in our task to evaluate the level of hallucination!

	MNIST	BraTS	MVTec AD
	Accuracy (%) (↑)	*Dice Coefficient* (↑)	*Accuracy (%)* (↑)
DDPM [11]	95.7±0.61	0.194±0.10	58.4±19.9
DDIM [32] (0.1T)	95.8±0.61	0.256±0.11	53.9±19.5
DDIM [32] (0.5T)	95.9±0.60	0.246±0.13	59.1±19.4
DDPM + DSI [38]	96.0±0.50	0.100±0.06	60.1±18.6
DDIM + Ours (0.1T)	96.1±0.30	0.447±0.10	66.8±17.8
DDIM + Ours (0.5T)	97.1±0.28	0.537±0.06	83.2±14.3
DDPM + Ours	**98.0±0.26**	**0.590±0.09**	**85.0±13.2**
Avg. gain	+1.5%	+0.293	+21.2%

+26%!

Tab. 2 Quantitative results on downstream tasks to measure hallucination



Main Results (Qualitative)

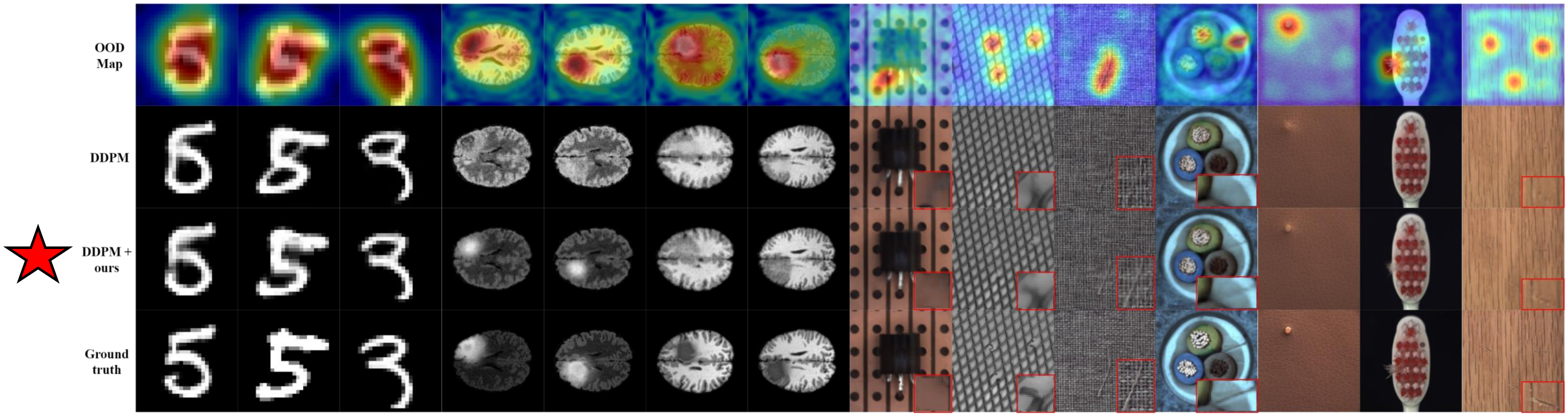


Fig. 10 Qualitative comparison on MNIST, BraTS and MVTec (From top: predicted OOD map, DDPM, DDPM with ours and ground truth).

Further Analysis

Does Local Diffusion work on various OOD?

-	Single DDPM	Single Ours	Multiple DDPM	Multiple Ours
PSNR (↑)	19.2	22.7	19.1	21.4
Dice coeff. (↑)	0.032	0.667	0.055	0.540

(a) Model's performance on single and multiple OOD in a single image

-	Small DDPM	Small Ours	Large DDPM	Large Ours
PSNR (↑)	20.7	23.4	18.0	21.5
Dice coeff. (↑)	0.050	0.685	0.030	0.580

(b) Model's performance on small ($< 1.5\%$) and large OOD ($> 3\%$) regions

Tab. 3 Quantitative comparisons on various types of OOD

Does Local Diffusion have negative impact on IND region?

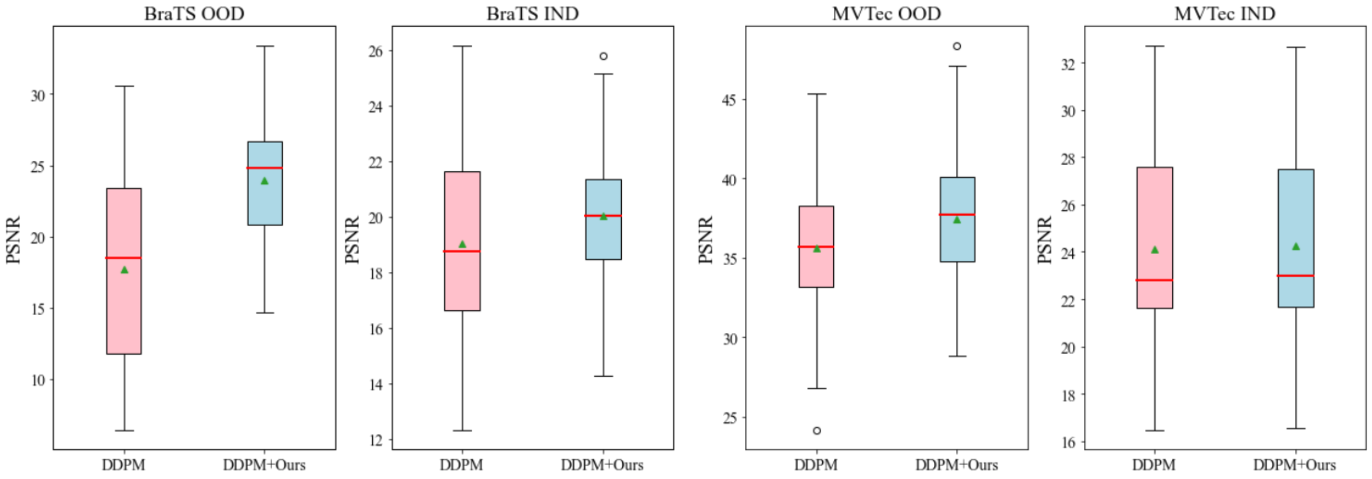


Fig. 11 Comparative analysis of performance across individual OOD/IND regions, The red lines and green dots represent the median and mean of each box, respectively

Thank you!

Arxiv **Github**