Harpoon: Generalised Manifold Guidance for Conditional Tabular Diffusion

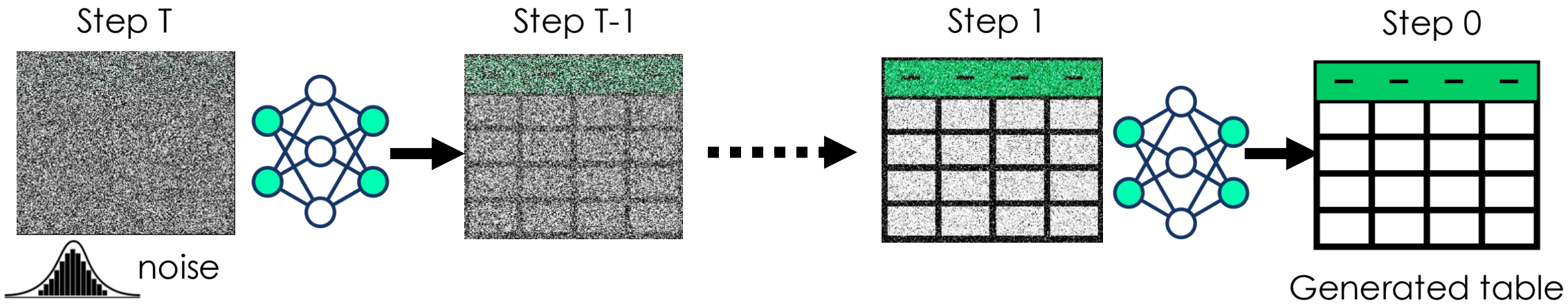
Aditya Shankar, Yuandou Wang, Rihan Hai, Lydia Chen

TU Delft Delft University of Technology

unine Université de Neuchâtel

Tabular Diffusion

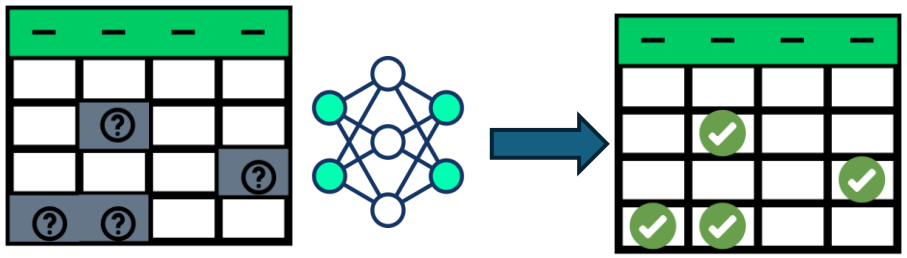
- Iterative generation from noise



Conditional Tabular Diffusion

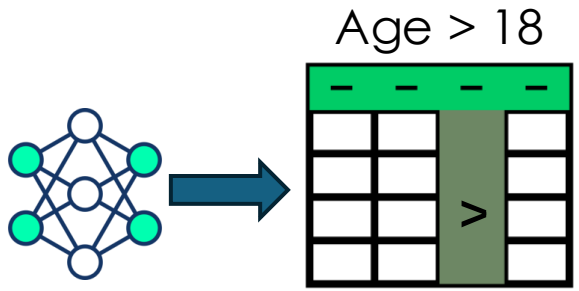
- Control the generative process

1 Imputation



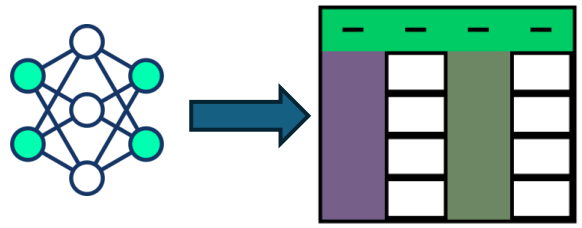
Imputation given partial observations

2 Inequalities



Inequality constraints: >, <, >=, <=

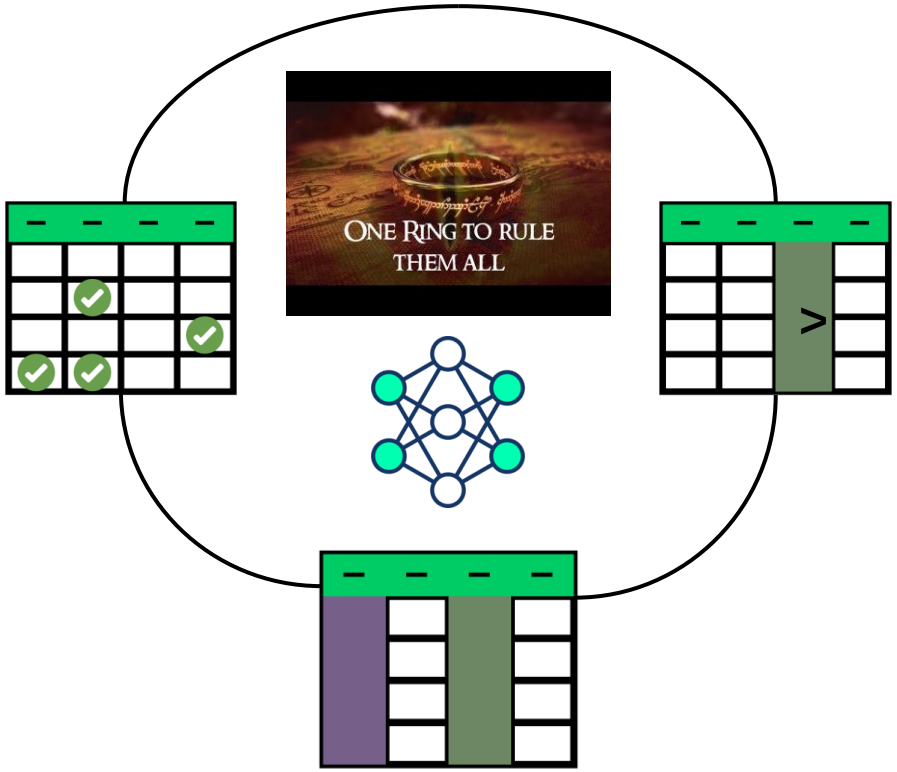
3 Combinations



e.g. Color='Blue' **AND** Age > 18

Challenges

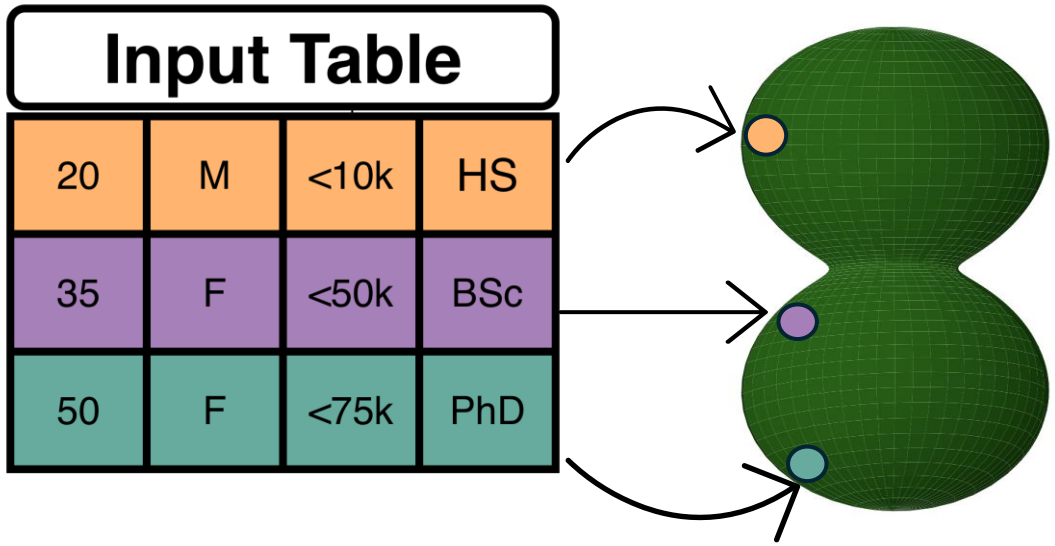
- One model for any task
  → Inference-time solution
- Theory-to-practice 
- Simple and efficient



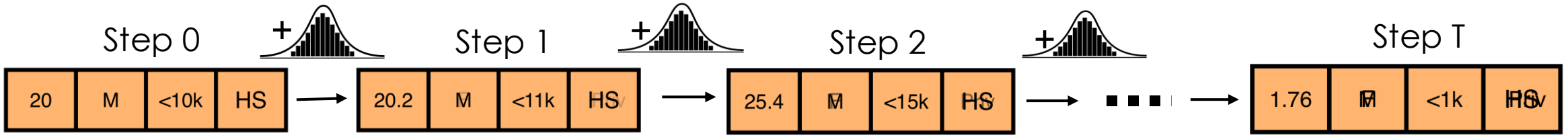
Harpoon

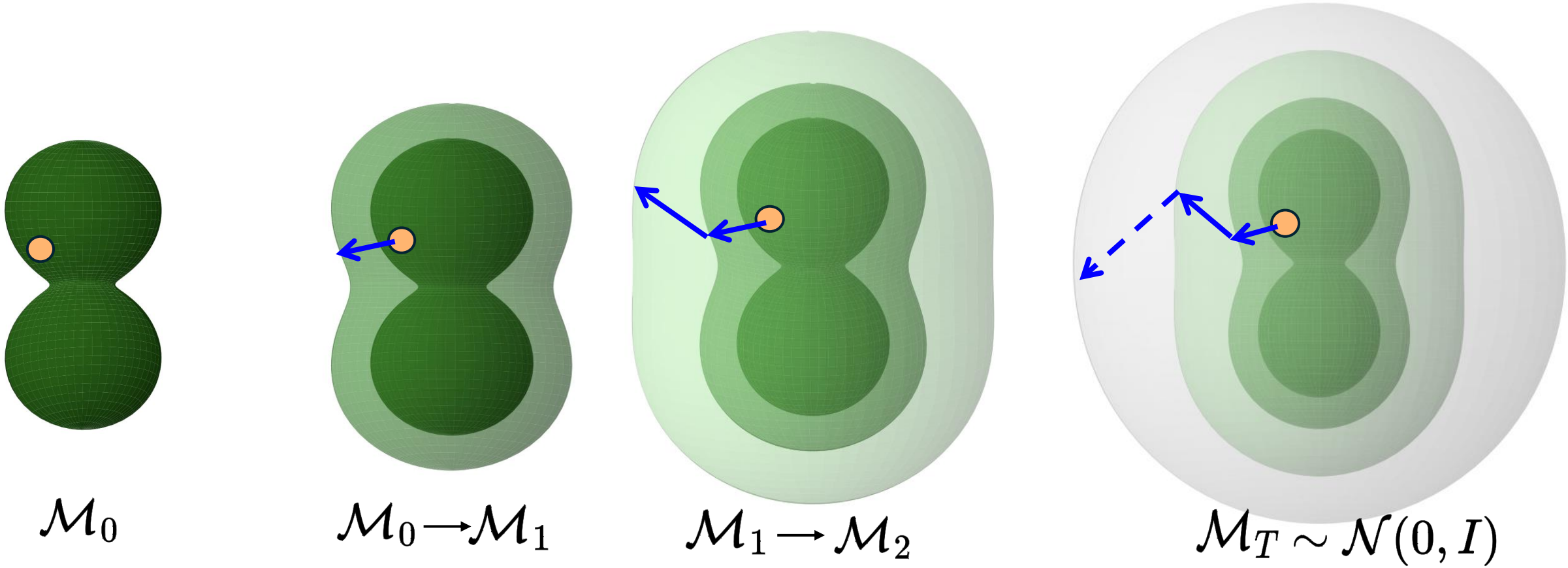
The Manifold Hypothesis

- Tabular data points assumed to lie on low-dim **manifold**



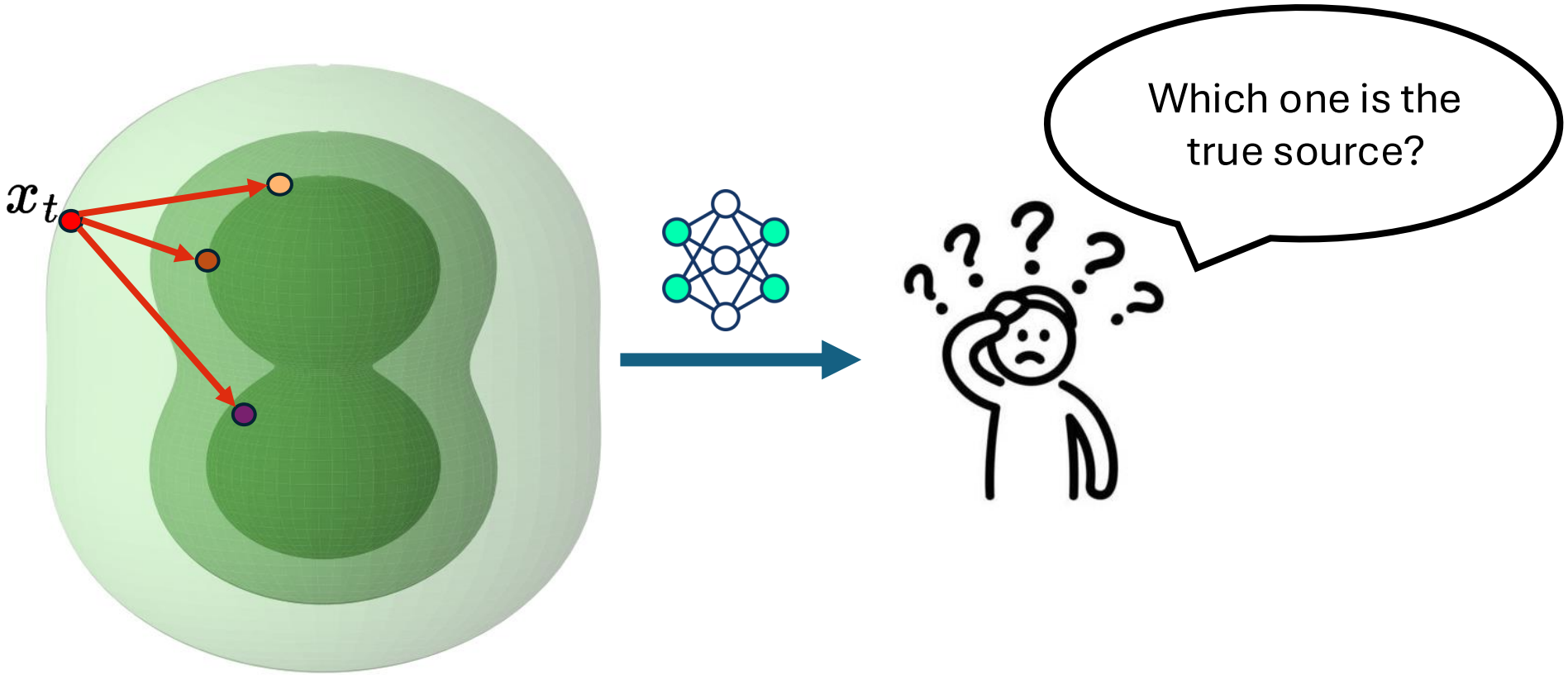
Geometric view of Tabular Diffusion





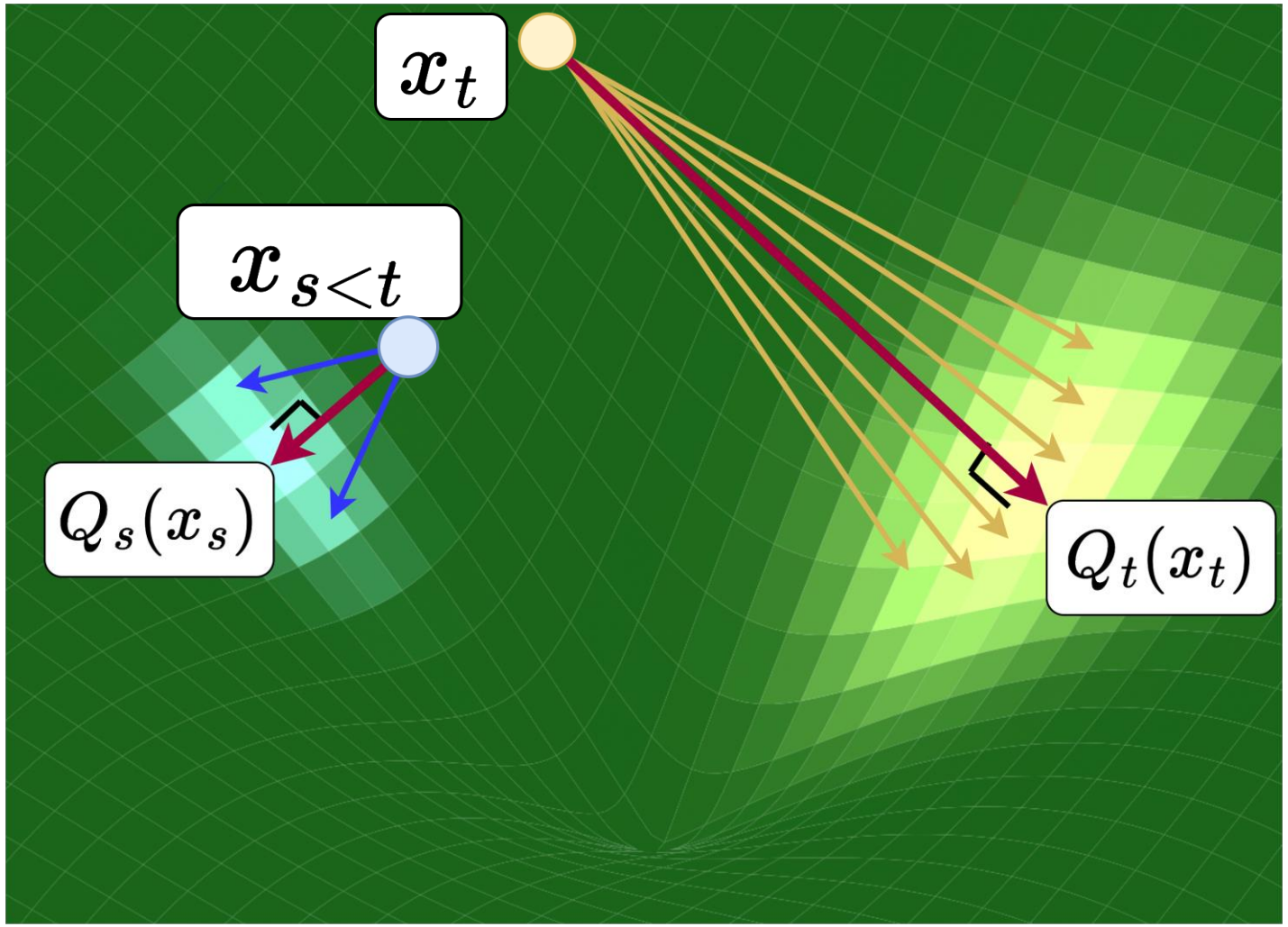
Diffusion model training

- For a given noisy point x_t
- Model learns to reverse added noise $\epsilon_\theta(x_t, t)$



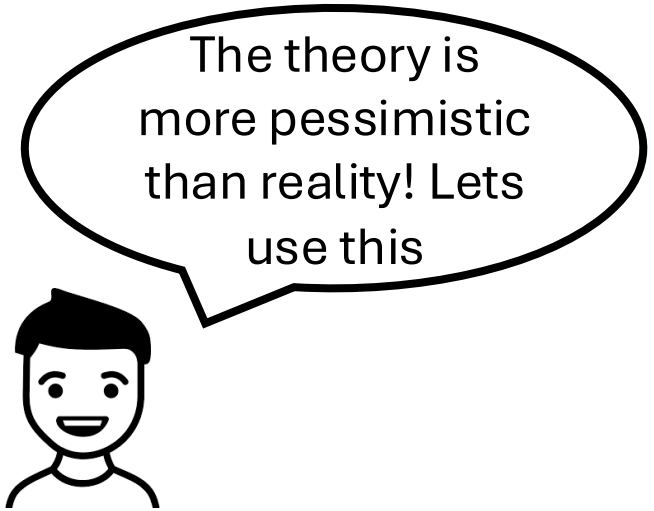
Limiting behaviour of model estimates

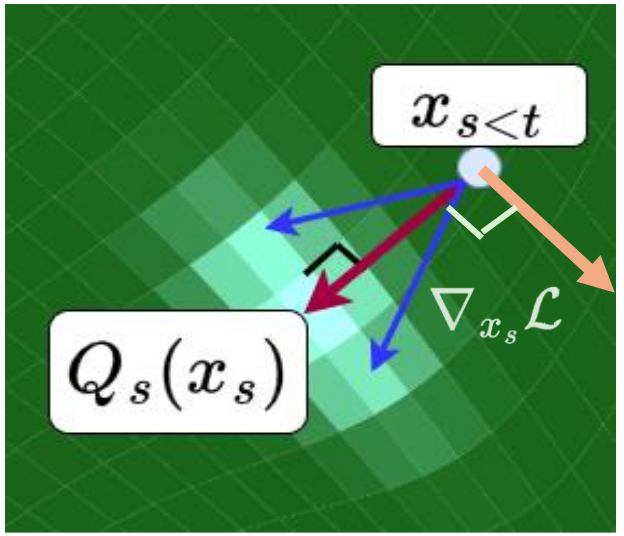
- Theorem 3.1 (Informal): "*At small noise levels, models converge to the* ***orthogonal projection*** *onto the manifold*"

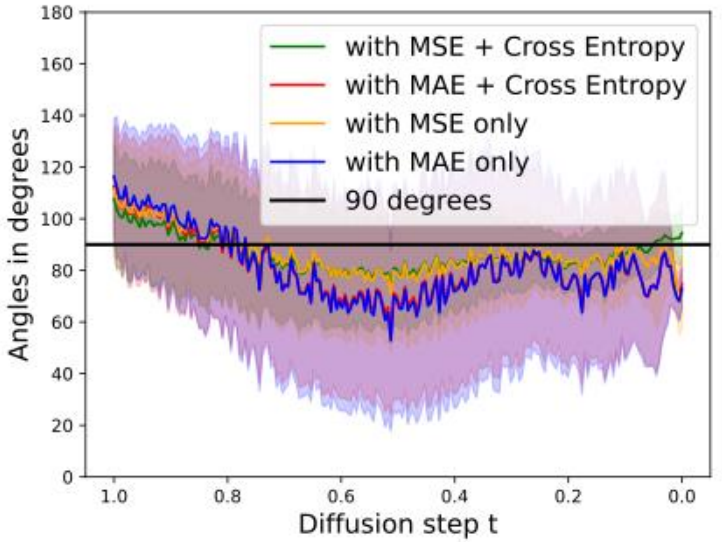


Limiting behaviour of gradients

- Theorem 3.2 (Informal): *"When the model projects orthogonally onto the manifold, the gradients from **any** differentiable loss w.r.t* x_t *become tangential to the manifold."*

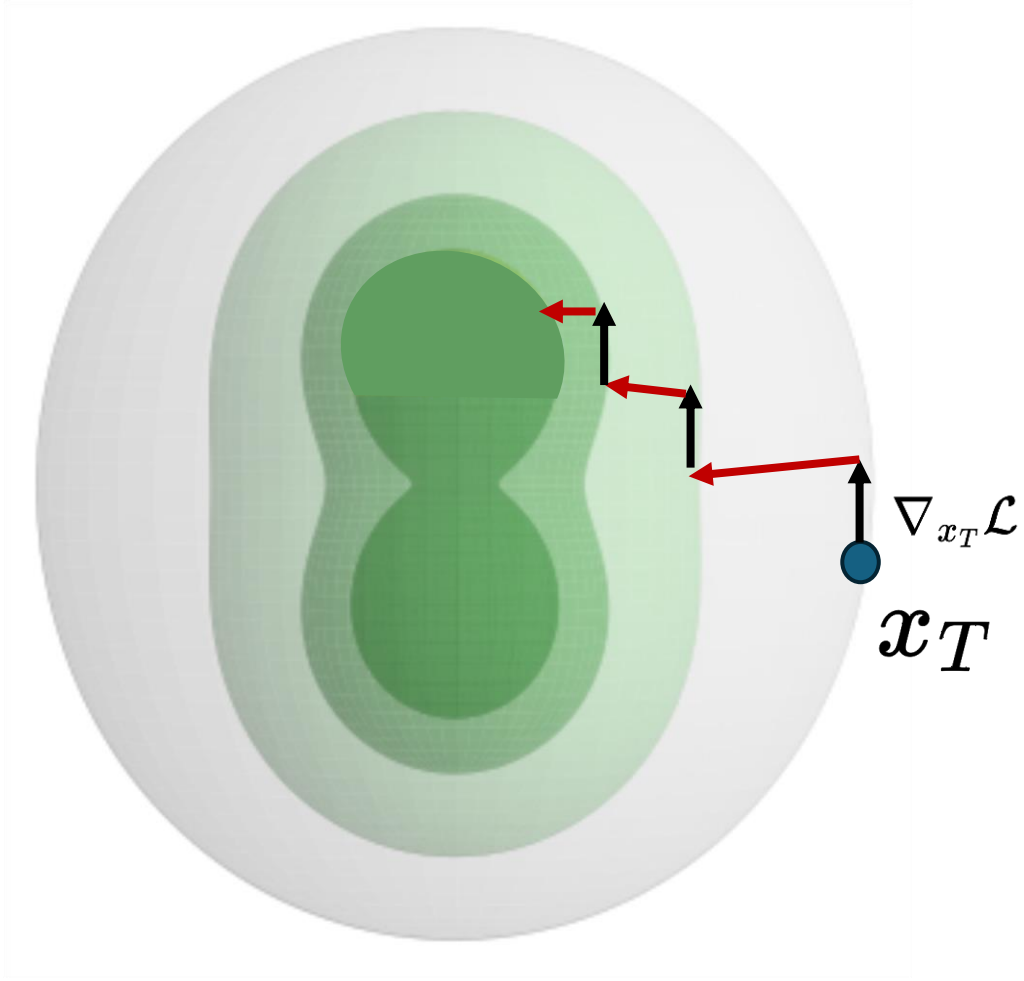




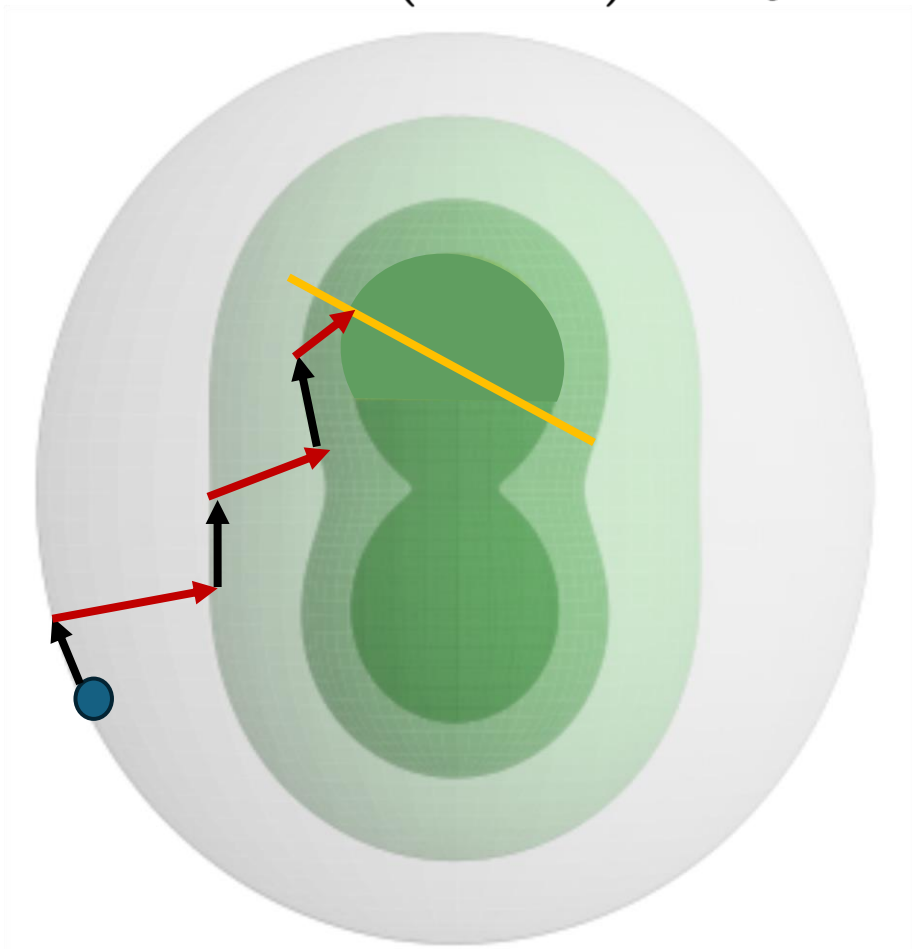


Algorithm: Inference-time guidance

e.g. 20 < Age < 25



Imputation: $(1-m)\odot x_0$

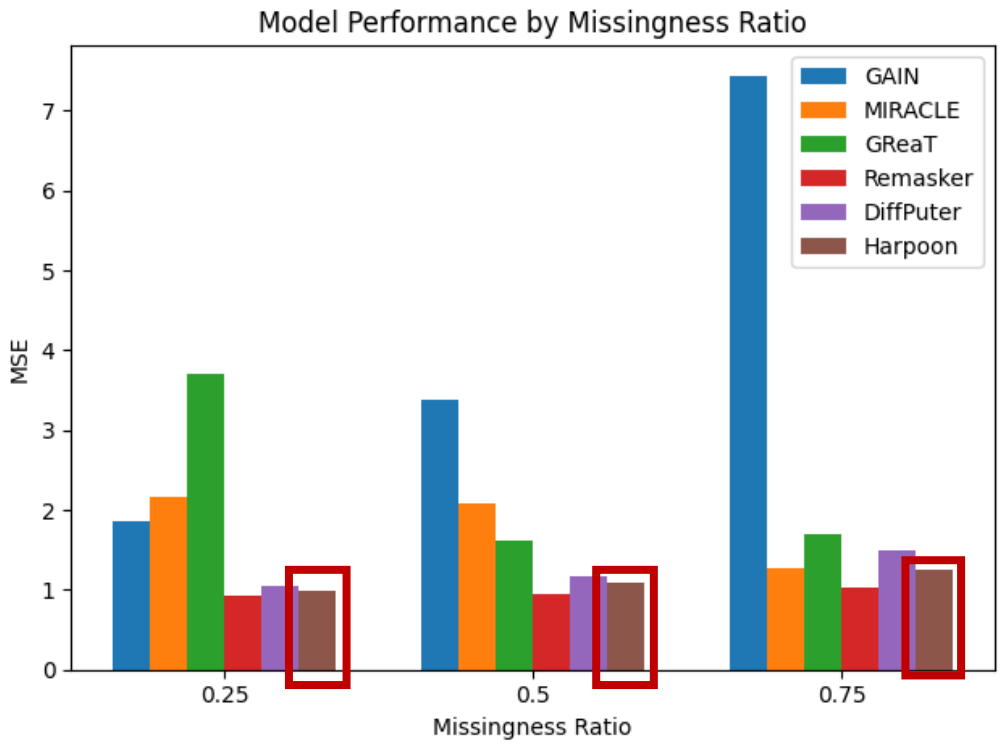


Experiments

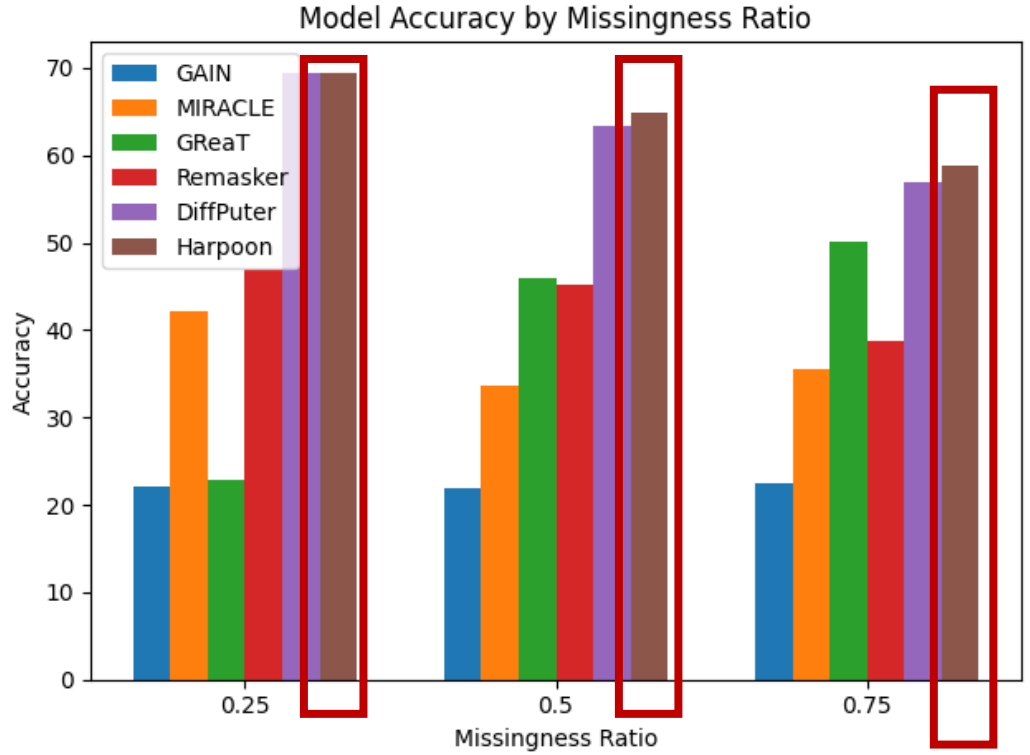
- 8 datasets
- Baselines: Diffusion, LLM, GAN, Causal Graphs, Transformer
- Imputation tasks
- Inequality constraints
- Multiple constraints (e.g. A **AND/OR** B)

Imputation Performance

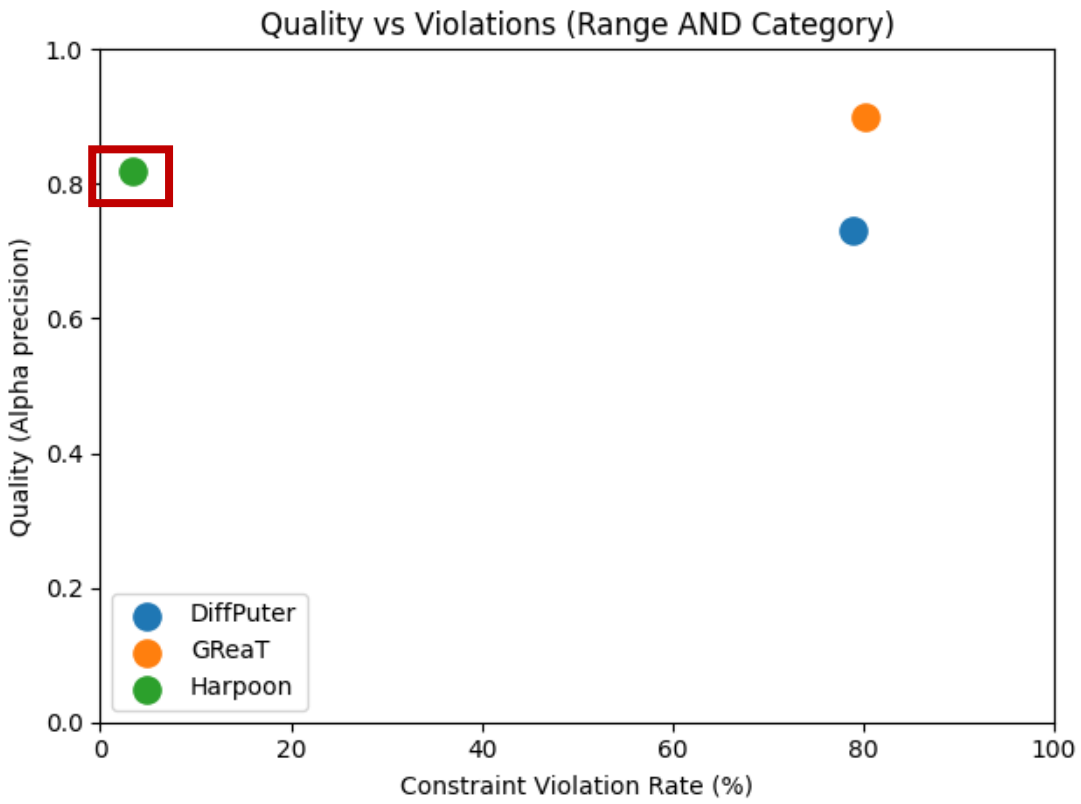
Continuous features (lower is better)

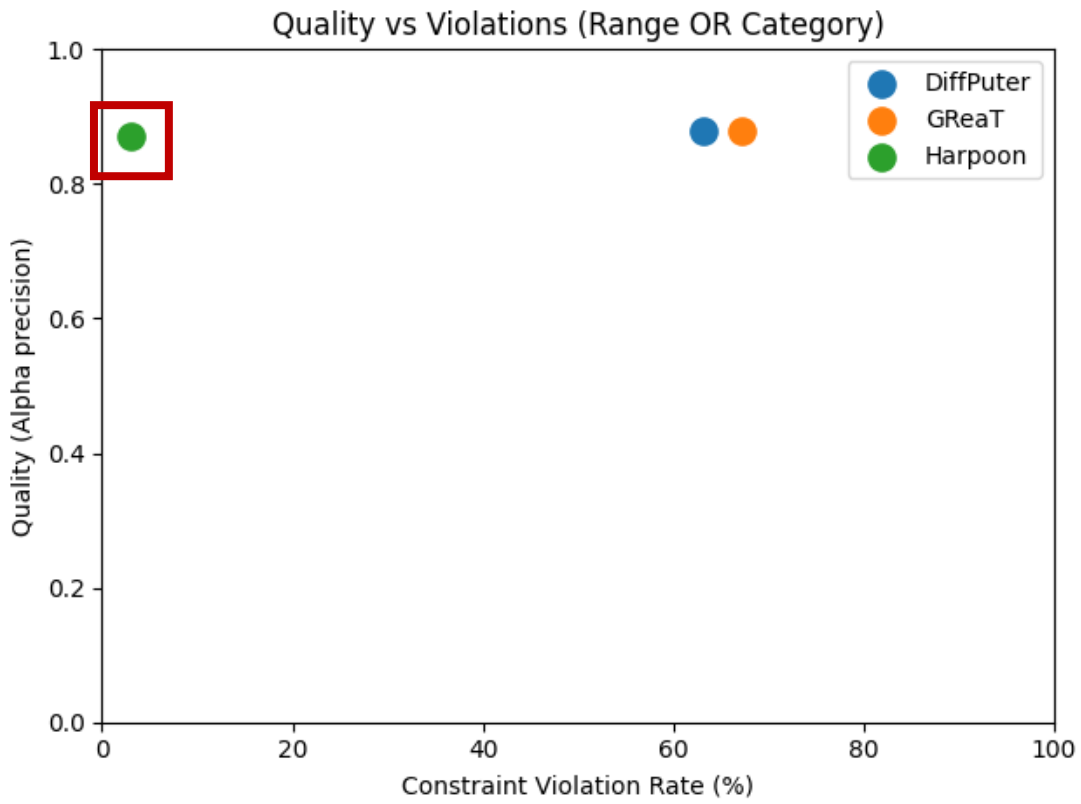


Categorical features (higher is better)



Multi-Constraint Generation Quality





To Conclude

- Harpoon is a controllable generator that is:
  - Theoretically grounded
  - Inference-time adaptable
  - Simple
- Future work:
  - Latent space control
  - Discrete noising mechanisms