Text-Conditioned Counterfactual CT Generation

Steering abdominal anatomy with report text



Stanford MEDICINE | School of Medicine

Current AI Evaluation Cannot Test Beyond Observed Data

Radiologist

- How would a disease appear if more severe?
- What if it involved a different lung region?
- What if disease progressed or regressed?
- What if this same patient had bilateral involvement?



Synthetic Data

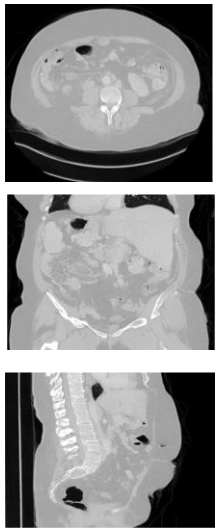
AI Models

- Evaluated only on observed images
- Cannot infer unseen severity variations
- Cannot infer unseen anatomical variations
- Failure modes remain hidden without explicit examples

Stanford

Synthetic Data Enables Controlled Clinical Variation

Real-World Data



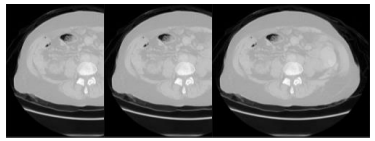
- No bowel obstruction
- Normal appendix
- No free fluid

Synthetic Data Generation

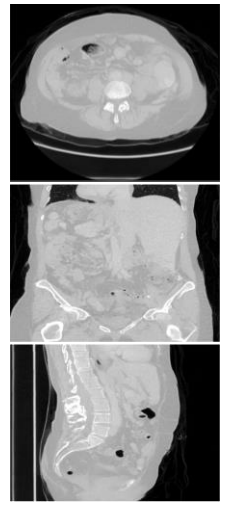
Edited Report Sentence

"Gastrointestinal tract: high-grade small bowel obstruction"

Text-Conditioned 3D CT Model



Real-World & Synthetic Data

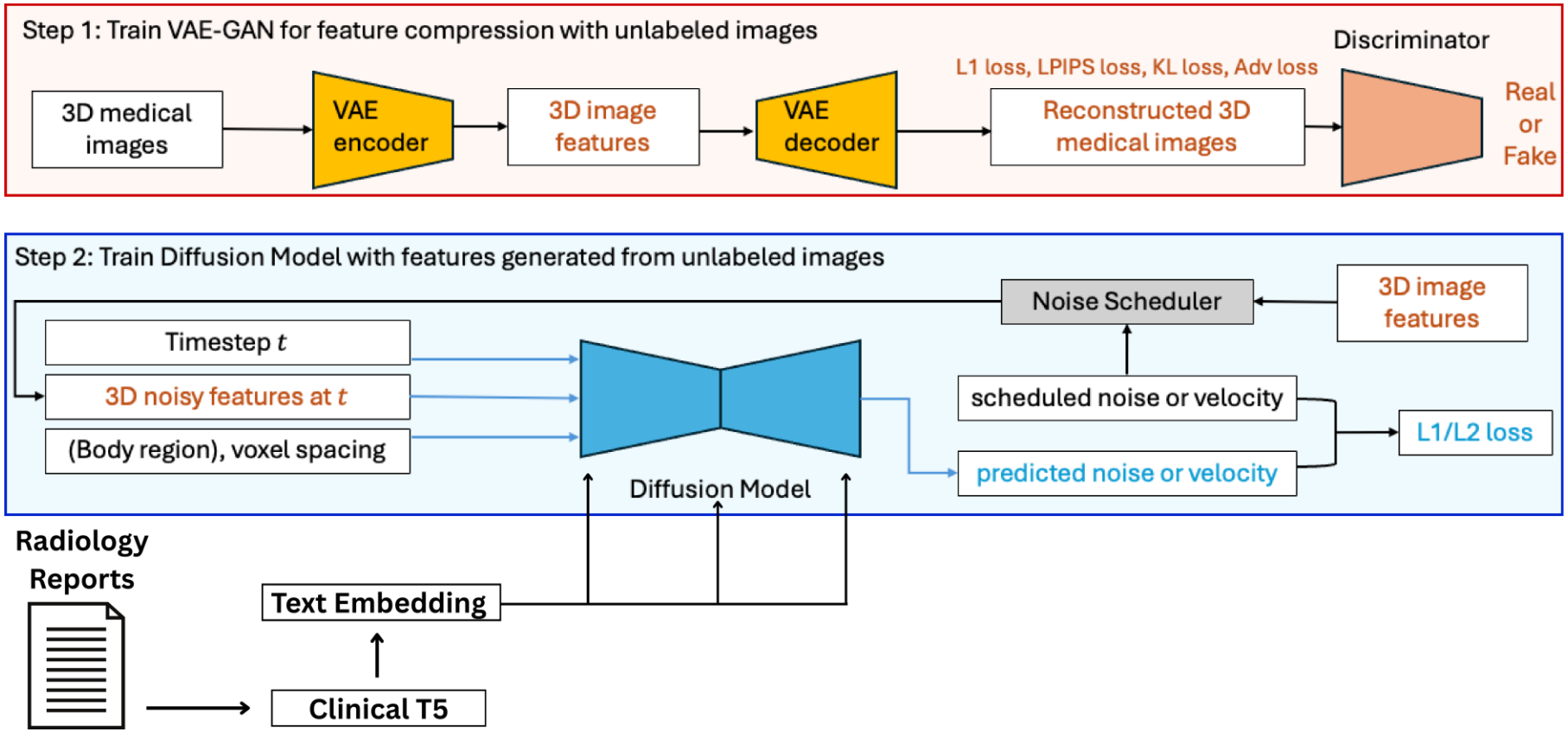


- High-grade bowel obstruction
- Moderate bilateral hydronephrosis
- Enlarged spleen
- Normal spleen

Same Patient → Controlled Variation

Stanford

Latent Diffusion for Text-Conditioned CT Generation



Backbone and Clinical-T5 frozen
only the text pathway is trained

Merlin: 25,489 CT + reports
pilot: 1,000 train / 100 val · 200 epochs

Stanford

Two Experiments

1



Which text representation controls generation?

impressions only · full report · LLM-structured

identical data, seeds, and hyperparameters — only the text differs

2



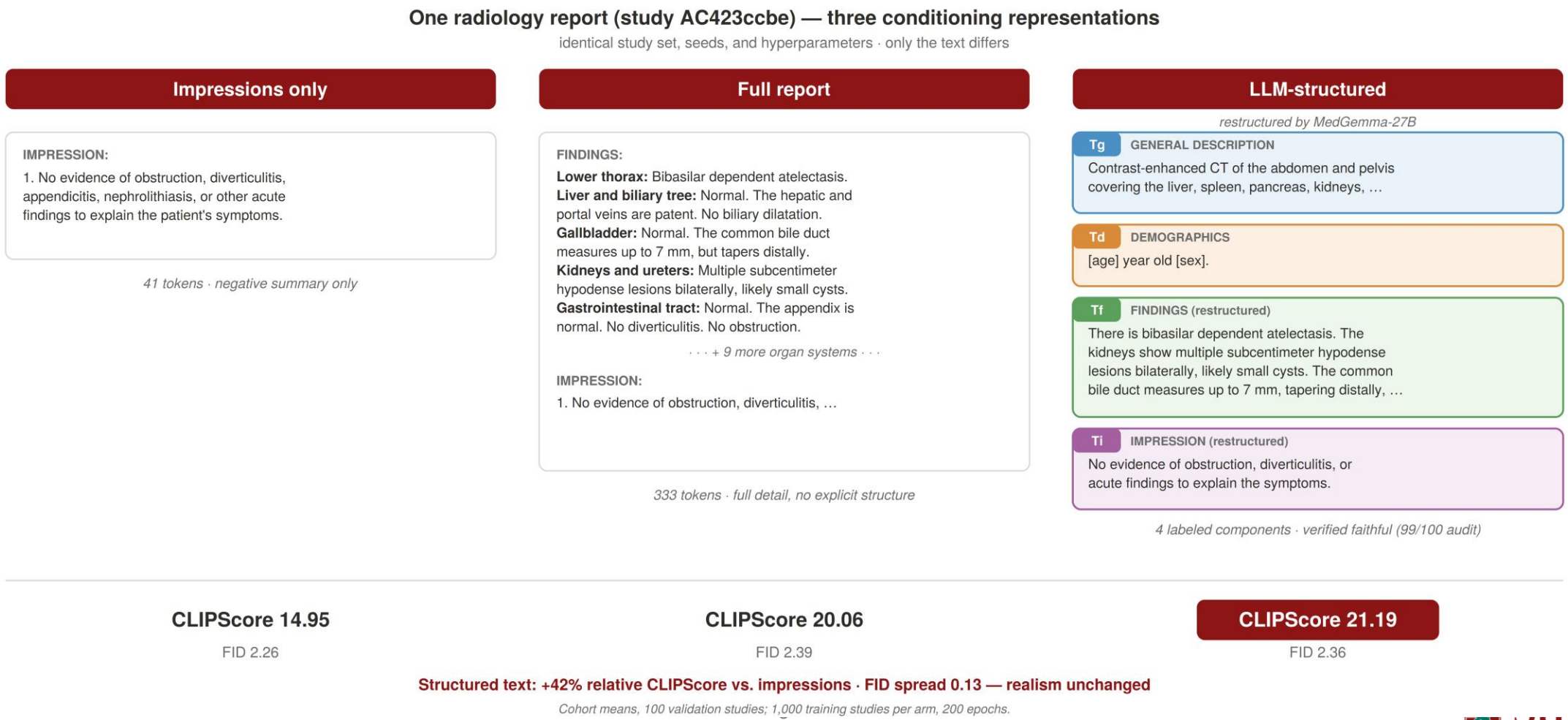
Can we generate counterfactuals?

single-sentence edits to the winning arm

fixed per-patient seed · SHA-256-verified identical initial noise

Stanford

One Report, Three Conditioning Representations



Stanford

Text Format Governs Alignment, Not Realism

+42%

relative CLIPScore

structured vs. impressions-only

FID spread: 0.13 across all arms
realism unchanged

Restructuring audited: 99/100 faithful
0 flipped negations · 0 invented findings

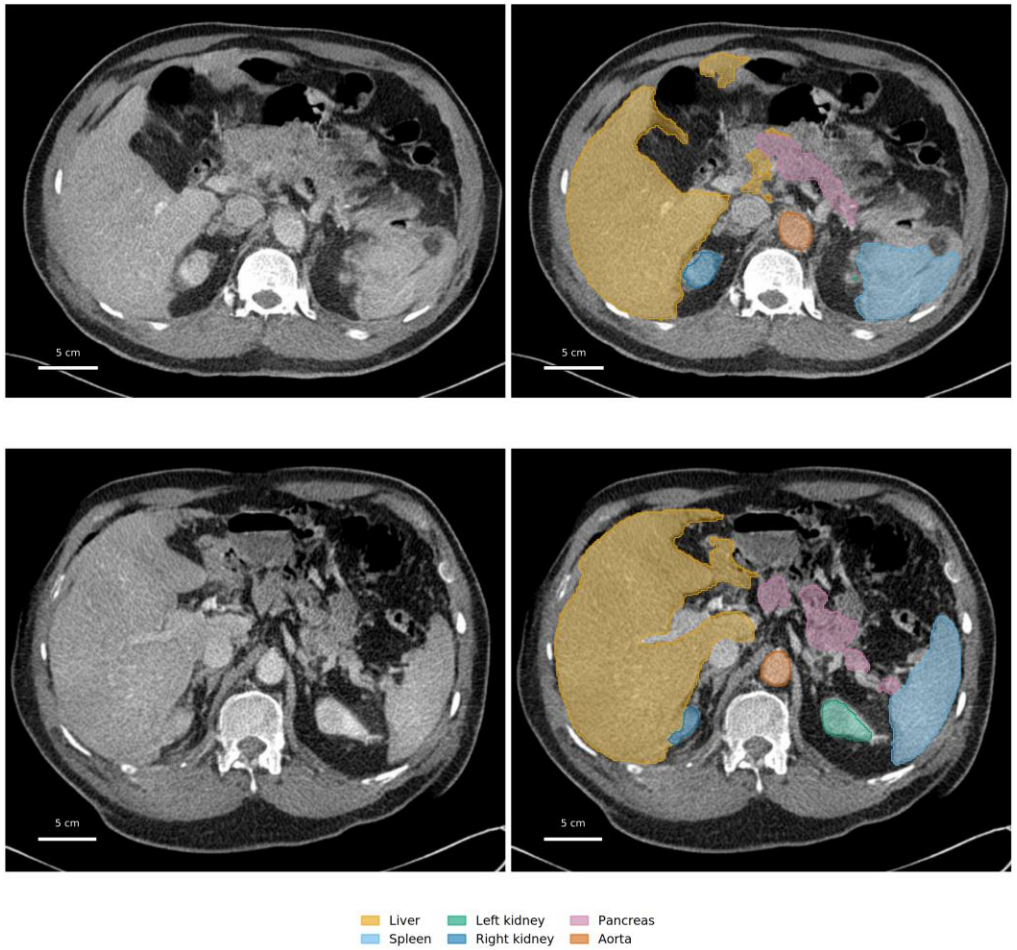
Conditioning arm	FID ↓	CLIPScore ↑
Impressions only	2.26	14.95
Full report	2.39	20.06
LLM-structured	**2.36**	**21.19**

All counterfactual experiments use the LLM-structured arm

Stanford

Generated Anatomy Is Present and Recognizable

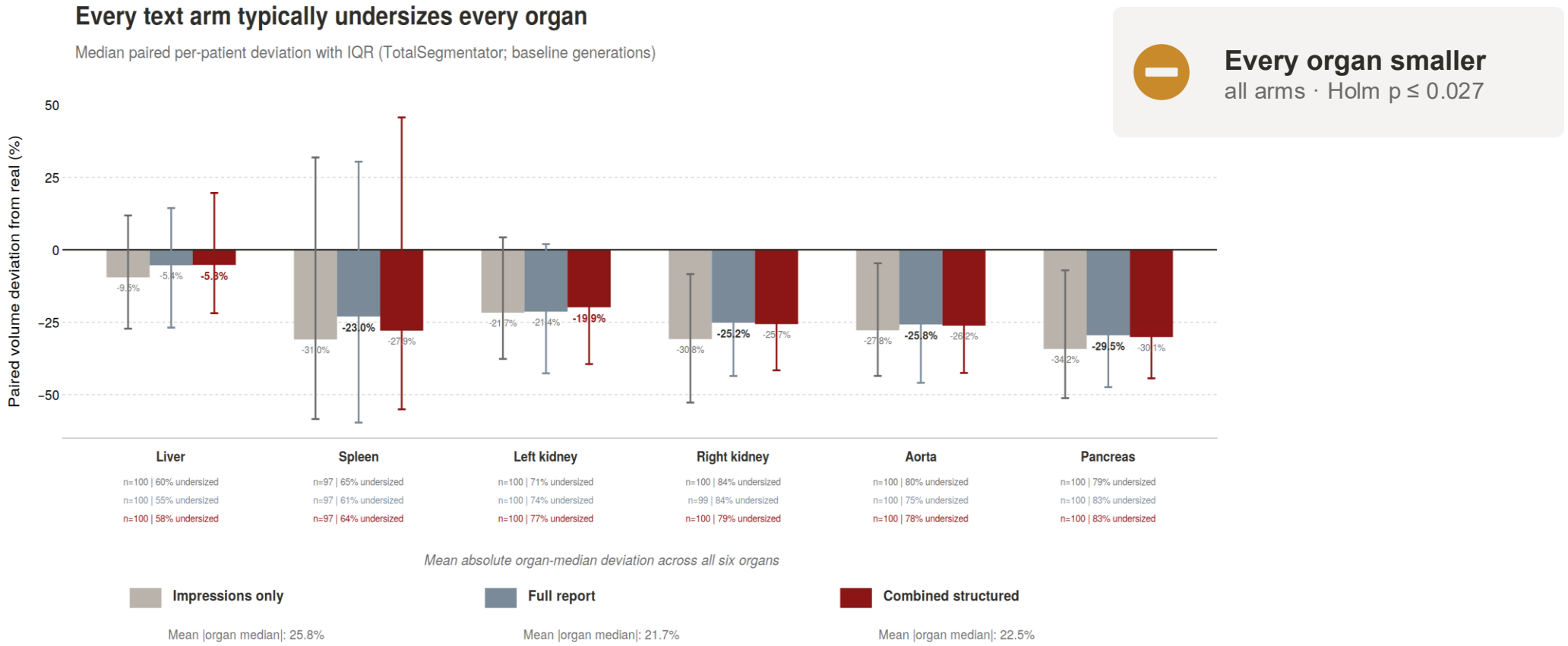
Two axial levels, same generated study



6/6 organs detected
in 100/100 generated volumes

Stanford

…but Systematically Smaller Than the Patient's Own Scan



Median-based ranking favors full report over combined structured; the cohort-means definition ranked combined structured first

Stanford

A One-Sentence Edit Produces a High-Grade Obstruction

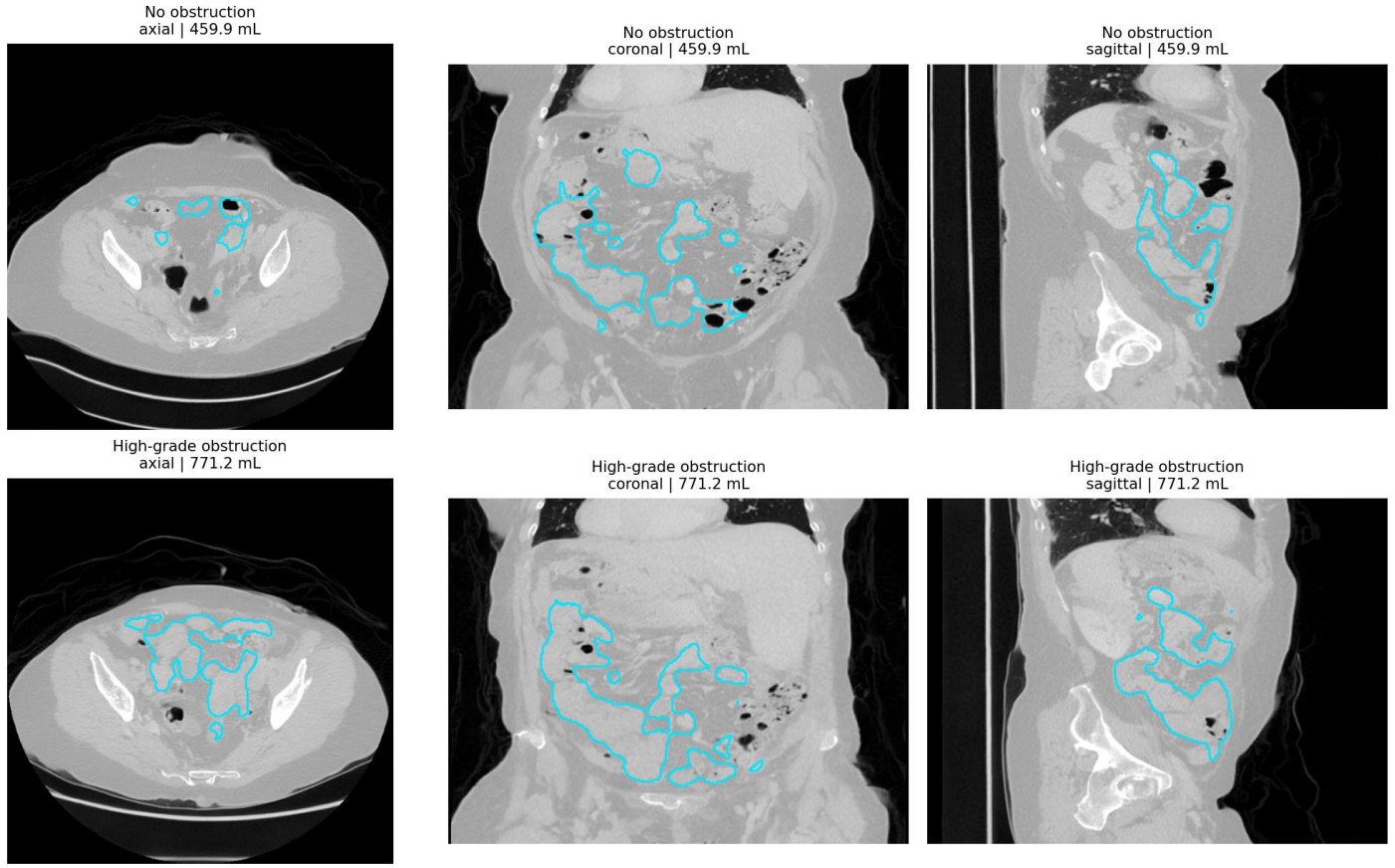
"High-grade small bowel obstruction with dilated loops measuring up to 4.5 cm"

one sentence edited · all else byte-identical · same initial noise

+311 mL

small-bowel volume · +68%

AC4215e71 | identical initial noise | primary change +311.3 mL



Stanford

Control Is Structural, Not Dimensional

Structural edit — obstruction

+68%

against the undersizing prior — real control

Dimensional — spleen ↓

−19.5 mL · p = 0.009

significant, but aligned with the prior

Dimensional — spleen ↑

p = 0.775

no effect against the prior

Stanford

Take-Home: Realism ≠ Clinical Utility



Text controls content, not realism

+42% CLIPScore · 0.13 FID spread — fidelity metrics cannot see controllability



Anatomy is average, not personal

organs always present, always undersized ($p \leq 0.027$)



Control is structural, not dimensional

obstruction succeeds; resizing fails — functional evaluation needed

Stanford

Limitations and Next Steps

Limitations

Pilot: 1,000 of 25,489 · single-run
effect sizes directional

Population-average anatomy

512-token truncation, alphabetical order
ρ = 0.974 · 7/23 spleen cases lost

Next steps

1. **Scale**
   full training · 12,245 structured studies (in progress)
2. **Focal vs. global panel**
   5 conditions · finding-specific readouts
3. **ControlNet masks**
   patient-specific anatomy
4. **Stress-test deployed AI**
   validated counterfactuals

Stanford

Thank You

Questions?

Rushil Srirambhatla
rsriram2@jhu.edu

Mentors: Magdalini Paschali · David B. Larson · Akshay S. Chaudhari
AIDE Lab · Department of Radiology · Stanford University School of Medicine

Acknowledgements

Stanford AIDE Lab for compute and feedback

Susu Sun for the ICH counterfactual framework

REU program organizers

Stanford