GRM-1245 A Synthetic Data Engine for Explainable Injection-Area Perception

Abstract

Vision-Language-Action (VLA) systems are beginning to support everyday clinical workflows. Deltoid intramuscular injection is a representative task, but progress is limited by data scarcity, privacy constraints, and the cost of expert annotation. Recent text-to-image (T2I) models make large-scale data synthesis possible, yet ensuring anatomical correctness, diversity, and label quality remains difficult.

To address this gap, we propose a Synthetic Data Engine tailored for medical perception, integrating cold-start filtering, controlled T2I generation, CLIP-based quality checks, and iterative segmentation training. We further introduce an anthropometry-grounded formulation of injection safety that produces interpretable safe-zone guidance. Experiments show that synthetic data can effectively bootstrap deltoid-segmentation performance and support reliable injection-area perception.

Introduction

As VLA technologies mature, robots and intelligent assistants are expected to operate safely in homes, clinics, and hospitals(Figure 1). In deltoid injection scenarios, workforce shortages and high cognitive load contribute to inconsistent technique and preventable nerve injuries. AI-assisted perception can provide stable recognition of anatomical regions, but progress is constrained by the lack of accessible, high-quality injection datasets.

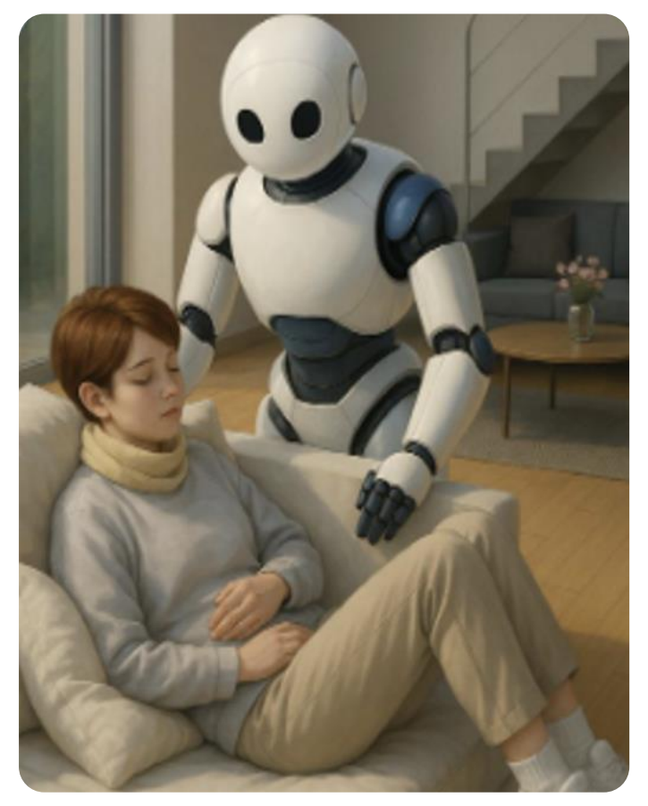


Figure 1. Robot in Family Scene

Real injection data are rare because of privacy restrictions, limited population coverage, and expensive expert labeling. While modern T2I generative models can create large numbers of images, reliably producing anatomically valid and clinically useful data remains challenging.

To address these gaps, we focus on robot-assisted deltoid injection and develop a unified Data Engine for the perception layer, with three contributions:

1. A modular, scalable synthetic-data pipeline(Data Engine) for rare medical perception tasks.
2. An explainable geometric framework\idea for estimating safe injection zones from perception outputs.
3. Dataset quality evaluation and a curated deltoid-segmentation dataset for downstream training and benchmarking.

Research Question(s)

1. Can a scalable Synthetic Data Engine accelerate training for rare, privacy-restricted medical perception tasks?
2. Can purely synthetic images achieve segmentation performance comparable to real-image training?
3. How can we provide explainable, anatomically grounded guidance for safe intramuscular injection based on perception outputs?

Materials and Preliminaries

CUDA Devices: Nvidia GPU 2080Ti*2 – 24 GB
Models: Yolo 11 Segmentation

Methods and Results

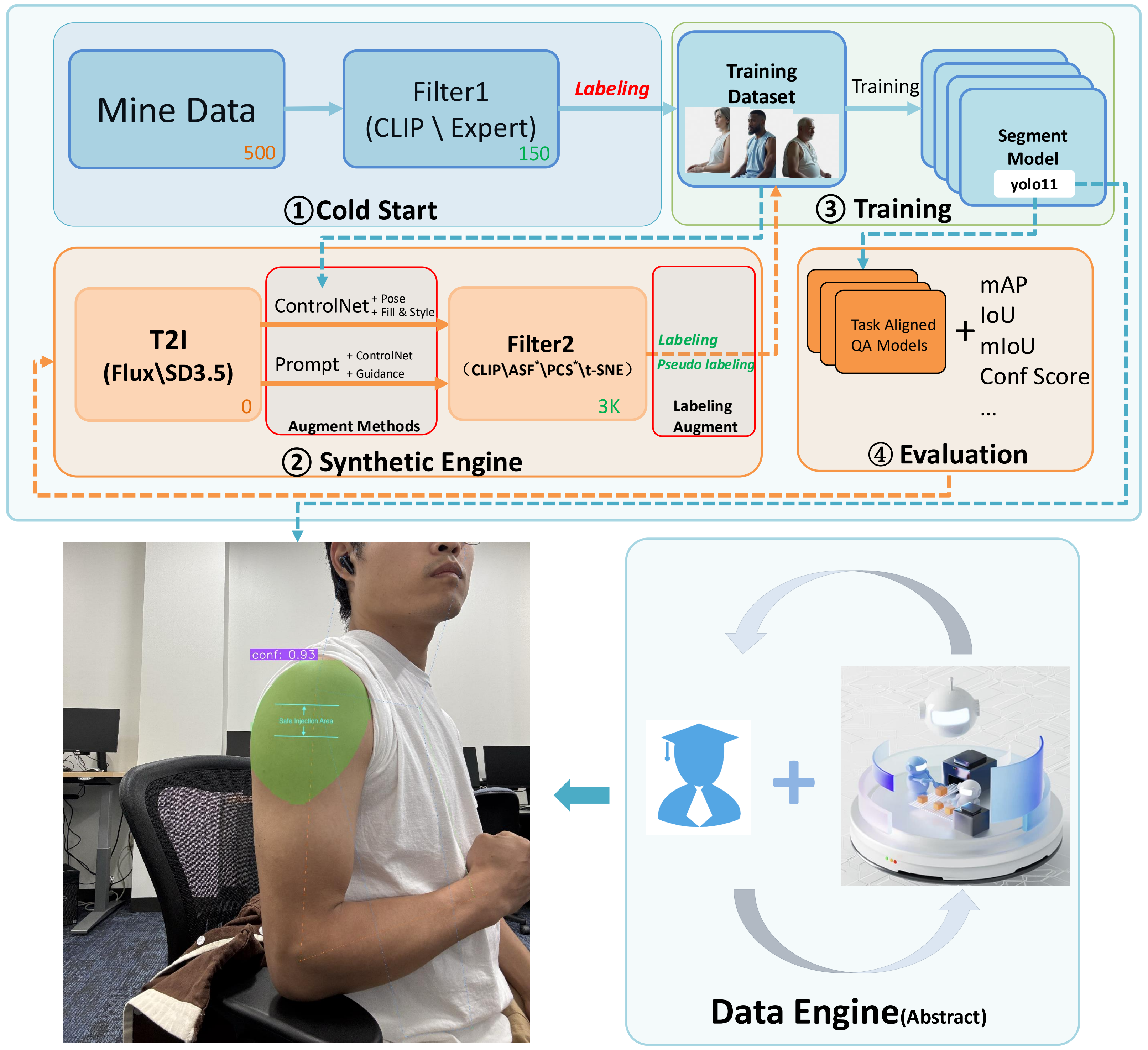


Figure 2. Overall architecture of the Synthetic Data Engine.

How to Enhance the Safety and Explainability for Medical Perception?

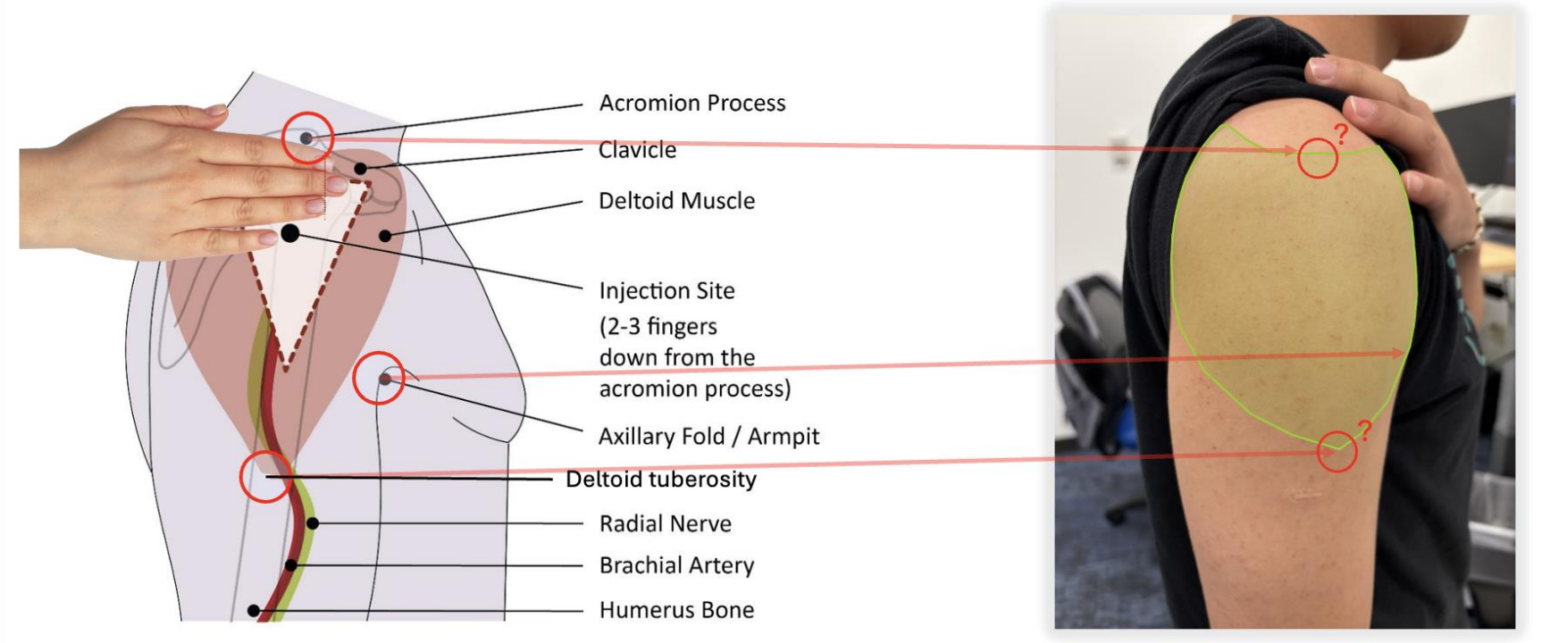


Figure 3. Anatomical structure of the deltoid injection site.

- **Anatomy**
- **Anthropometry**

$$
\begin{cases}
L = \|A - E\| \\
AP = \langle P - A,\ \dfrac{E - A}{\|E - A\|} \rangle^{*} \\
\mathrm{Safe}(P) = \begin{cases} 1, & 0.20L \le |AP| \le 0.30L \\ 0, & \text{otherwise} \end{cases}
\end{cases}
$$

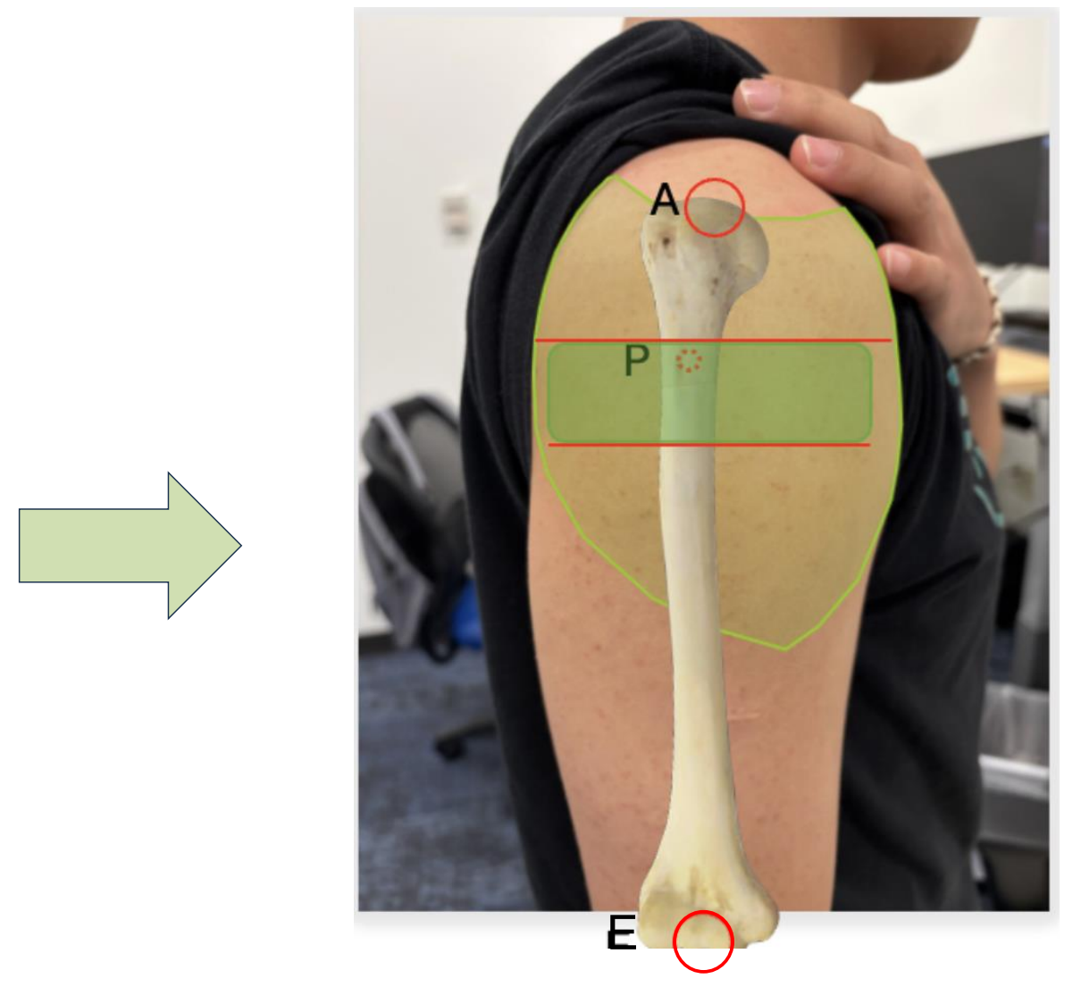
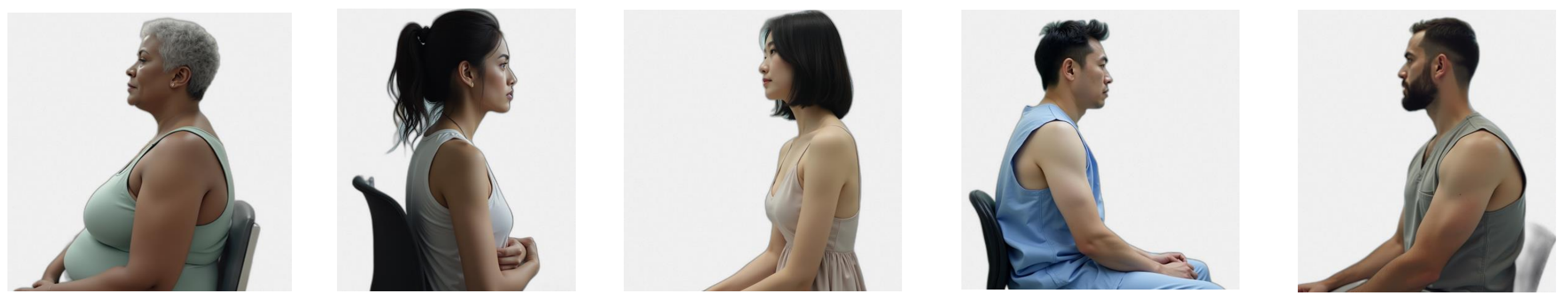


Figure 4. Explainable injection zone with anatomy- and anthropometry-based auxiliary lines.

Our High-quality Synthetic Dataset（3K）



Task QA Model for Injection Task

Guardrails

Recall ≥ 0.90

Precision ≥ 0.90

F1 ≥ 0.90

Metrics Sensitivity

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{AP_i^{\mathrm{real}} - AP_i^{\mathrm{synthetic}}}{AP_i^{\mathrm{real}}}\right| \le 8\% \ ^{*}$$

$$\mathrm{dist}_{\%} = 100\% \cdot \frac{\Delta c}{d_{gt}} \le 10\%$$

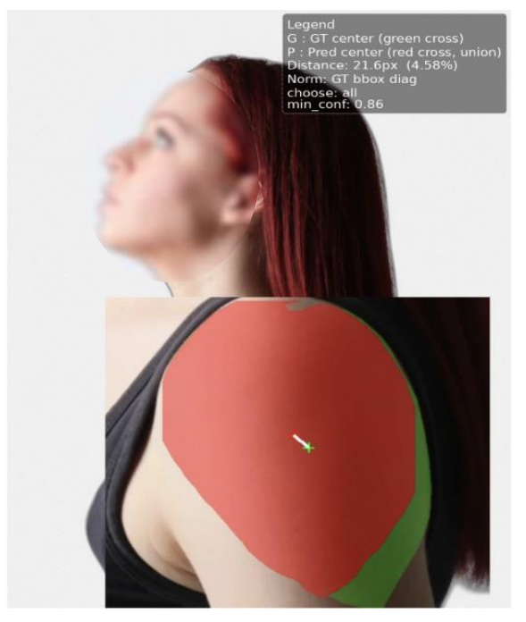


Figure 5. *Predicted vs. ground-truth injection centers on real data.*

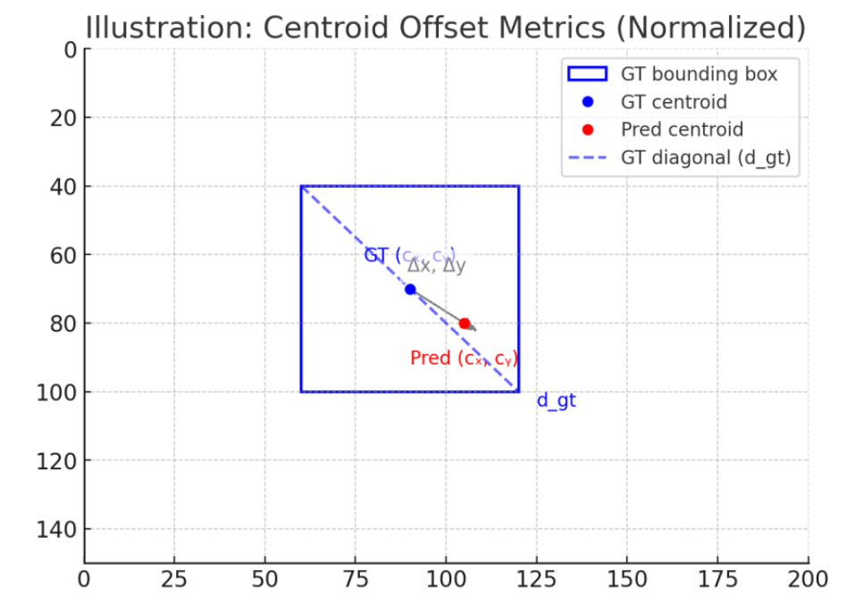


Figure 6. *Illustration of normalized centroid offset.*

Core QA Evaluation Result

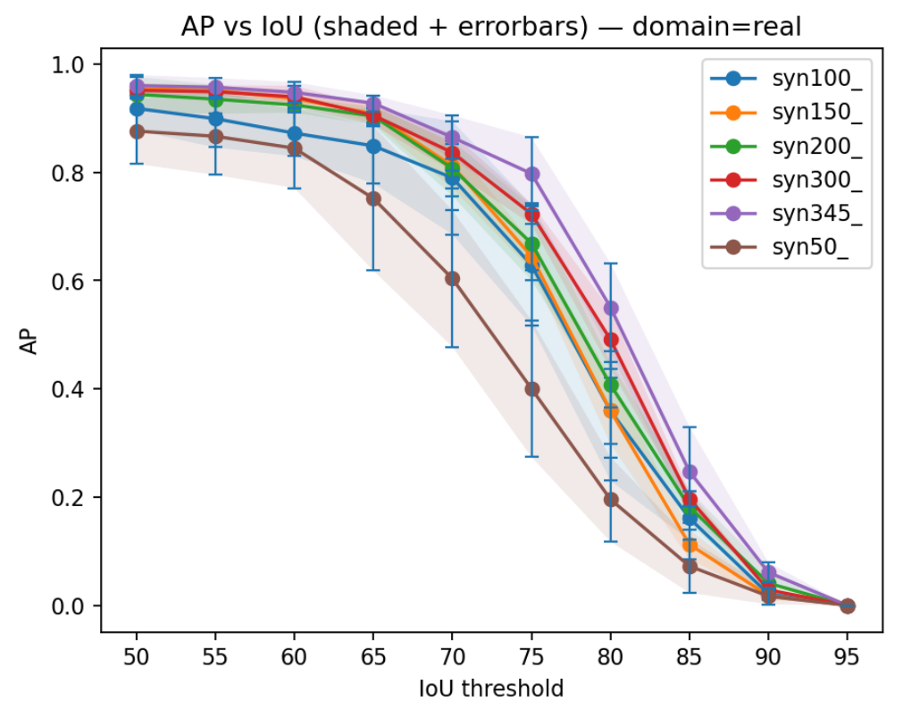


Figure 7. Performance across synthetic-only datasets of varying sizes (50–345).

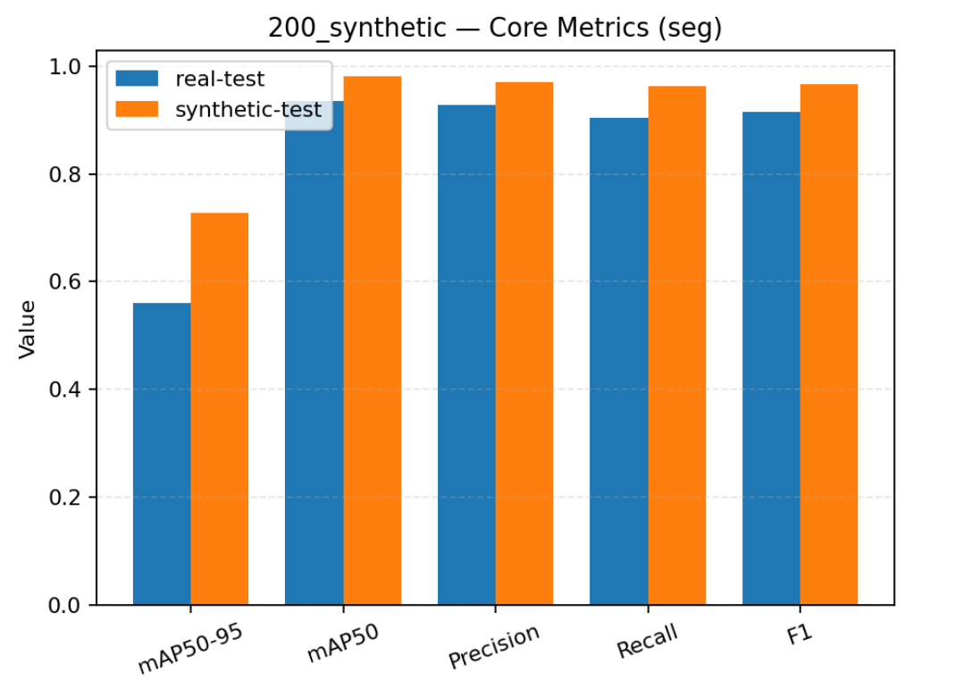


Figure 8. Core metrics on the 200-image synthetic dataset.

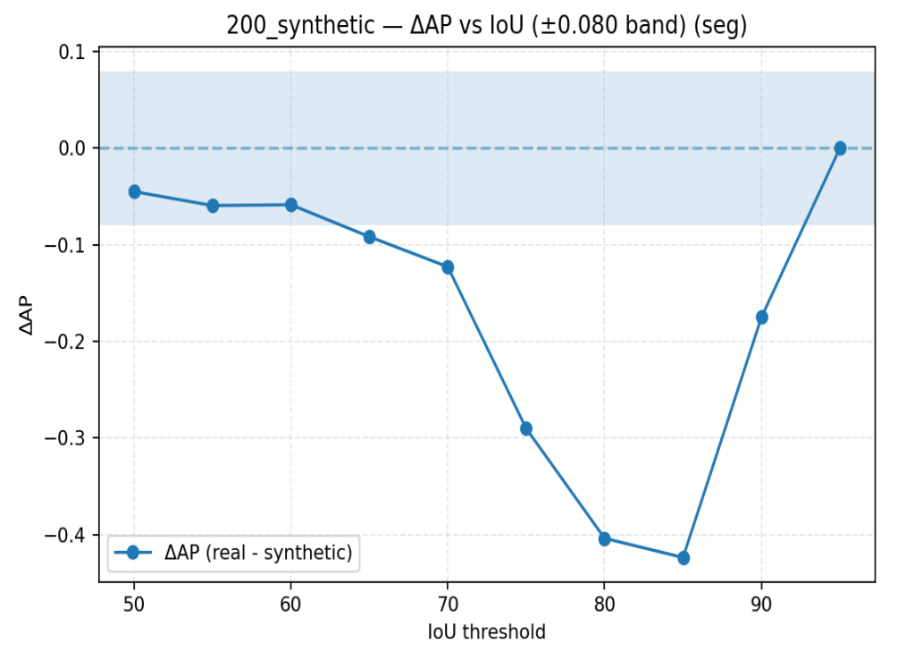


Figure 9. Gap between real-test and synthetic-test evaluation.

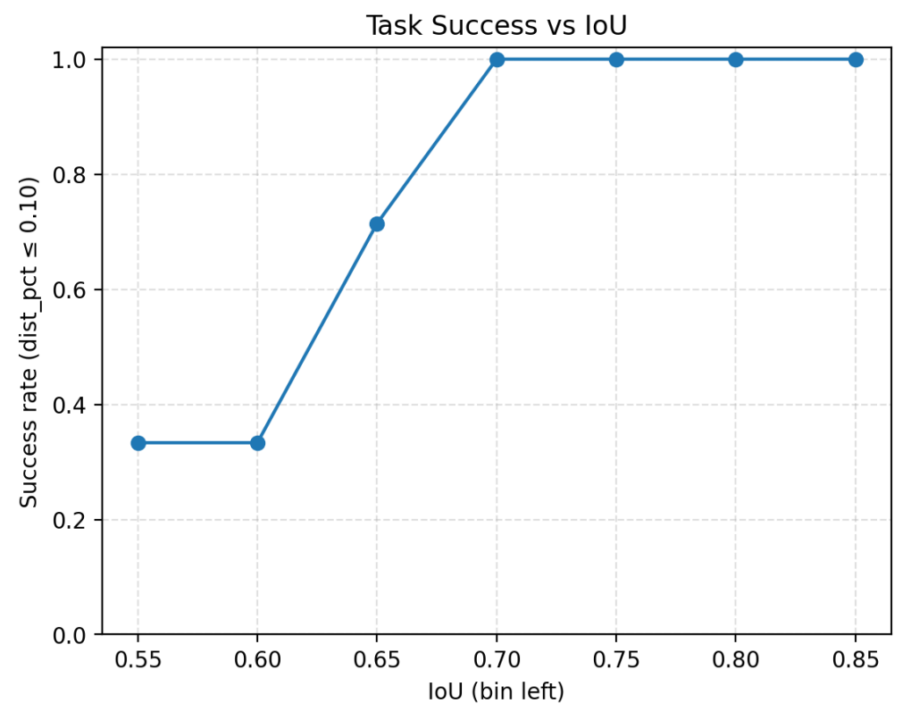


Figure 10. Relationship between task-success rate and IoU.

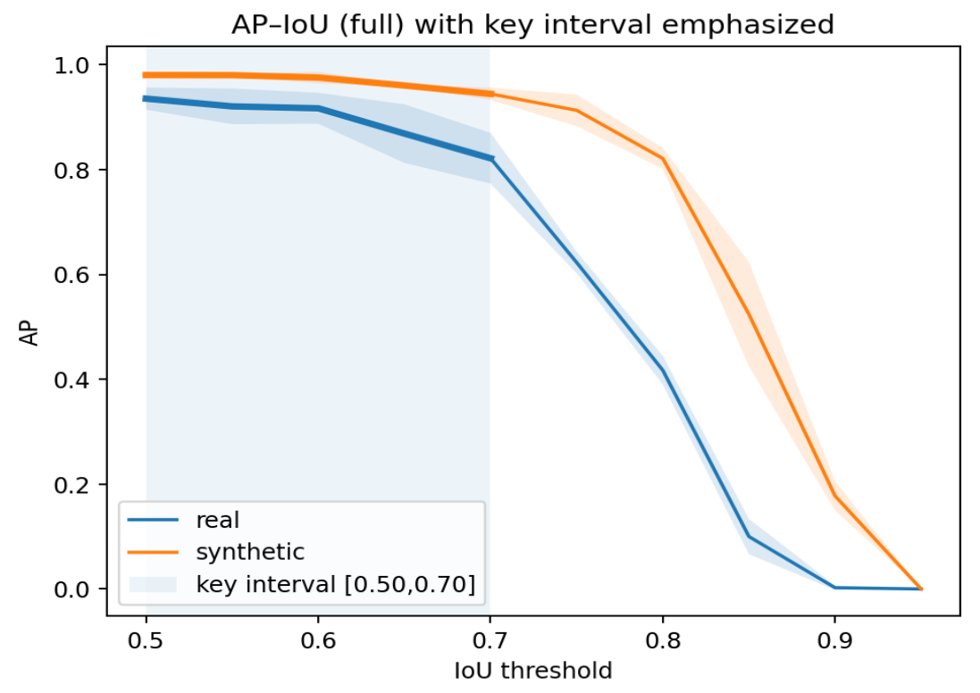


Figure 11. Domain-gap visualization via AP–IoU curves for real vs. synthetic gap evaluations.

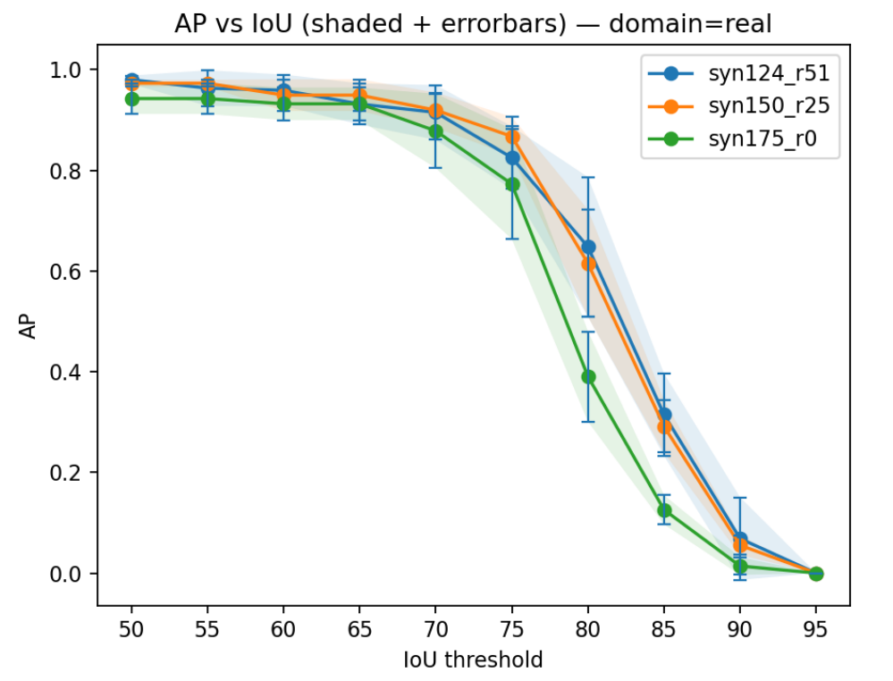


Figure 12. Performance effect of mixed synthetic-real training.

Conclusions and Feature Work

Our Synthetic Data Engine achieves real-level segmentation performance, with synthetic-only training **reaching ≥90% AP** of real data and staying **within MAPE ≤ 8%.** The QA model ensures reliable localization **(dist% ≤ 10%)** and mixed synthetic–real training yields further performance gains.

The anthropometry-driven framework provides **consistent and interpretable injection-zone safety guidance.**

We plan to finalize and open-source the full Data Engine pipeline, advance 3D-aware perception for higher accuracy and spatial consistency, integrate the system into a broader VLA framework for real-world medical and robotic applications.

Contact Information

yshen4@students.kennesaw.edu

References

1. The Recommended Deltoid Intramuscular Injection Sites in Adults (Charmode et al., Cureus 2024)
2. Upper Limb Muscle Volumes in Adult Subjects (Holzbaur et al., J. Biomech 2007)
3. A Training-free Synthetic Data Selection Method for Semantic Segmentation(Tang et al., arXiv 2025)
4. SAM 3: Segment Anything with Concepts (Anonymous, **ICLR** 2025, under review)
5. CLIP is Also an Efficient Segmenter: A Text-Driven Approach for Weakly Supervised Semantic Segmentation(Lin et al., **CVPR** 2023)
6. FreeMask: Synthetic Images with Dense Annotations Make Stronger Segmentation Models (Yang et al., **NeurIPS** 2023)
7. Identifying Cases of SIRVA in the United States: NLP-based Detection Method (Zheng et al., JMIR Public Health Surveill 2022)

➢ **Notes** *

L: Upper-arm length. **A:** Acromion point. **E:** Elbow point. **AP:** A proportional position along the A-E axis (not a geometric projection).

ASF: Annotation similar filter (based on annotation class counts and CLIP similarity). **PCS:** Perturbation CLIP similar filter.

Δc: represents the Euclidean distance between the predicted injection-center point and the ground-truth center.

MAPE: average percentage error between real and synthetic AP values across IoU thresholds.

KENNESAW STATE
UNIVERSITY
COLLEGE OF COMPUTING AND
SOFTWARE ENGINEERING

Author: Yukang Shen
Advisor: Dr. Yan Huang