



Western



Dataset Access

# Synthetic Data Alone is Enough? Rethinking Data Scarcity in Pediatric Rare Disease Recognition

Ganlin Feng<sup>1</sup>, Yuxi Long<sup>1</sup>, Erin Lou<sup>2</sup>, Lianghong Chen<sup>1</sup>, Pingzhao Hu<sup>1,2,\*</sup>, Wei Xu<sup>2,\*</sup>  
<sup>1</sup>Western University, <sup>2</sup>University of Toronto | \*Correspondence: phu49@uwo.ca, wei.xu@uhn.ca

CVPR  
JUNE 3-7, 2026

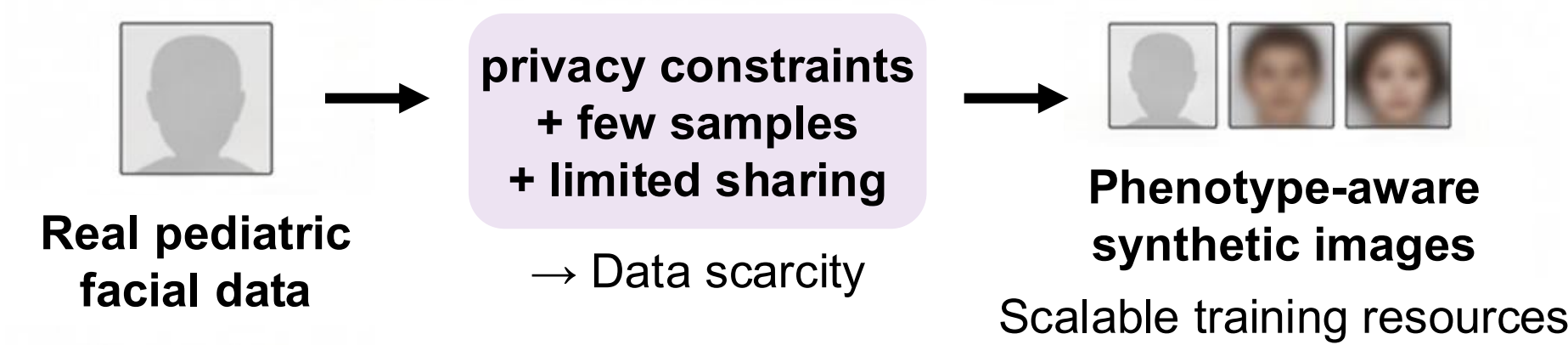


DENVER  
COLORADO

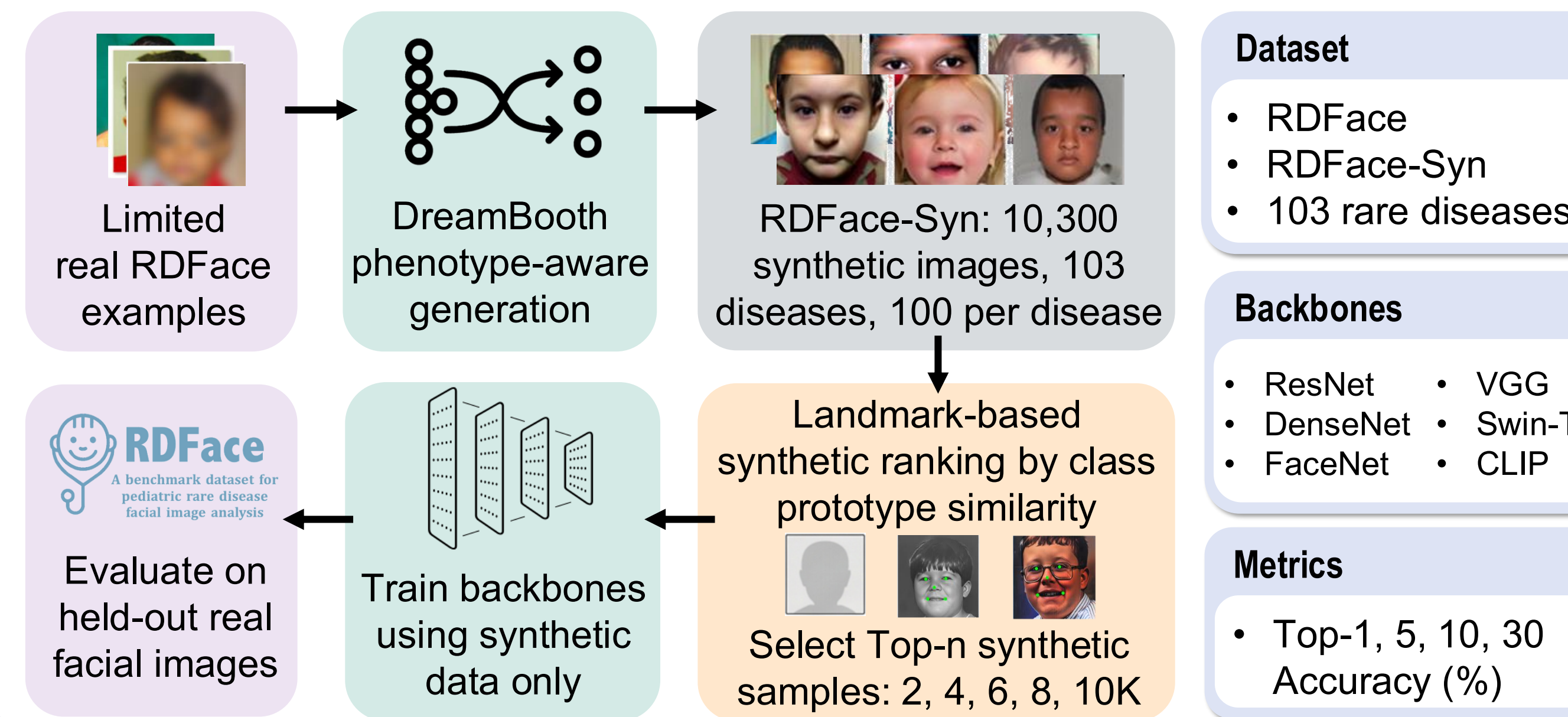
CV4CHL Workshop

## Why Synthetic-Only learning?

- Pediatric rare disease recognition is limited by extreme data scarcity
- Children's facial images are privacy-sensitive and difficult to share
- Real clinical datasets often contain only a few samples per disease
- Synthetic images may provide scalable, shareable visual resources for model training and education

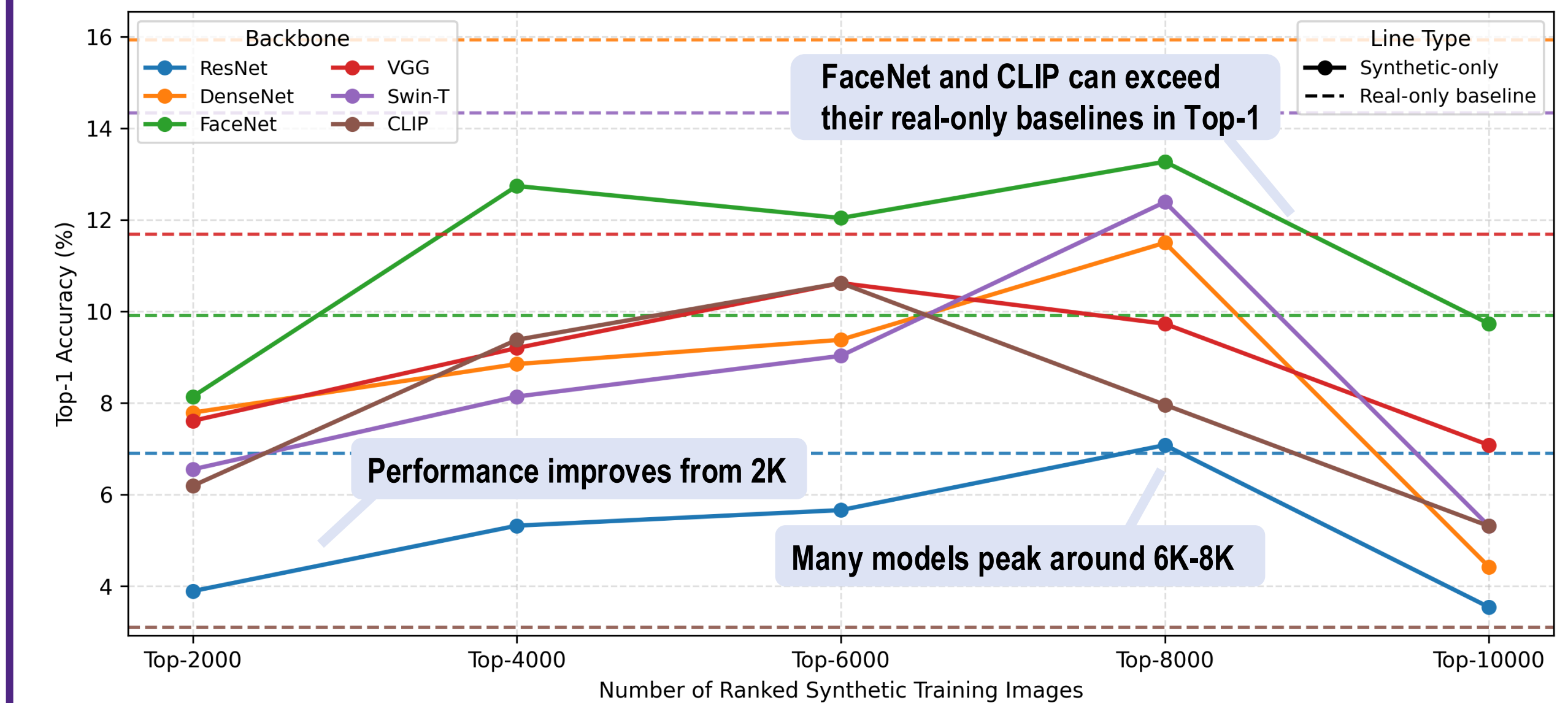


## Synthetic-Only Evaluation



## Key Findings

### Top-1 Accuracy vs. Synthetic Data Scale

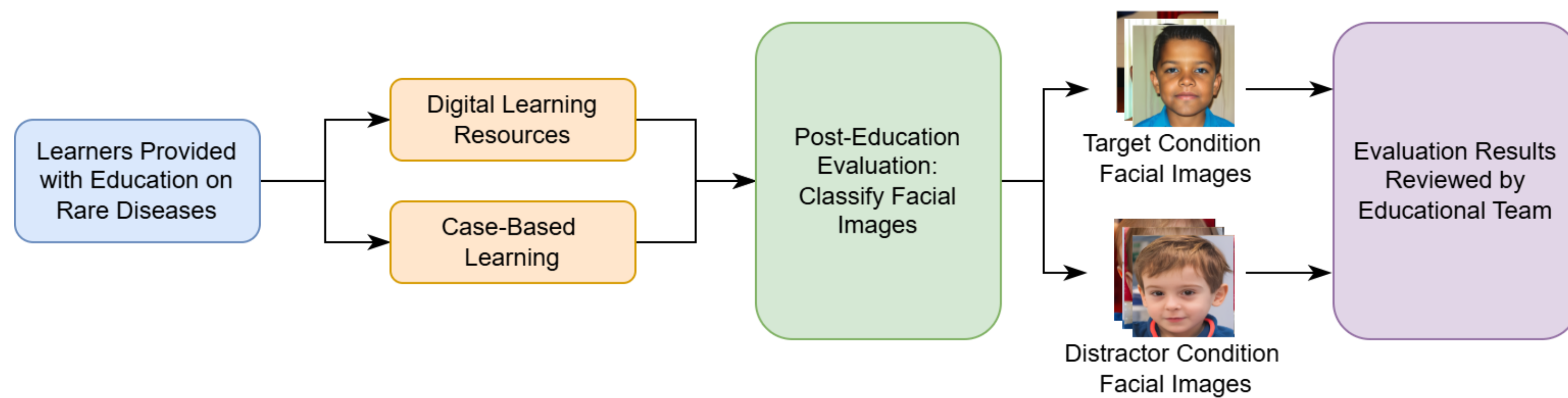


**Main Takeaway:** Synthetic-only training can approach real-data-only performance at sufficient scale, but effectiveness depends on backbone architecture and synthetic sample quality.

- Synthetic-only models become competitive at intermediate synthetic scales.
- Best performance often occurs around 6K–8K samples.
- More synthetic data is not always better: lower-ranked samples may introduce noise.

## Clinical and Educational Value

### Beyond Classification: Privacy-Preserving Pediatric Education



- Reduces reliance on identifiable pediatric facial images
- Supports genetic education and clinician training
- Provides shareable condition-specific visual reference
- **Future work:** improve synthetic fidelity and validate with clinical user studies

Supported by:



Chaires de recherche du Canada

Canada Research Chairs



Edwin S.H. Leong Centre for Healthy Children UNIVERSITY OF TORONTO



UNIVERSITY OF TORONTO

