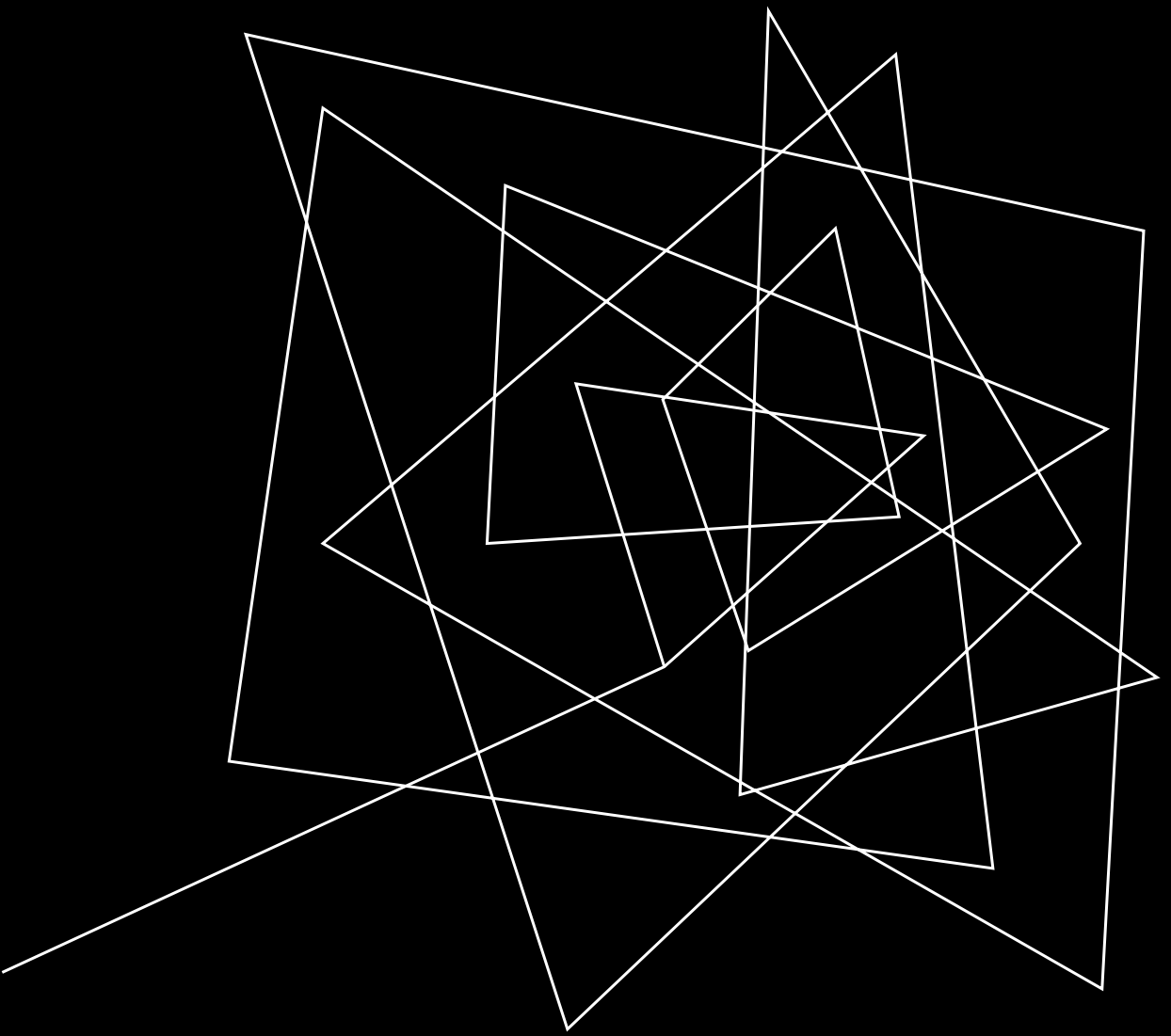


MULTI-HUMAN TESTBENCH – BENCHMARKING IMAGE GENERATION FOR MULTIPLE HUMANS

Presenter: **Shubhankar Borse**
Staff ML Researcher, Qualcomm AI Research

Paper: https://arxiv.org/abs/2506.20879
Code/Data: https://github.com/Qualcomm-AI-research/MultiHuman-Testbench/tree/main

AGENDA

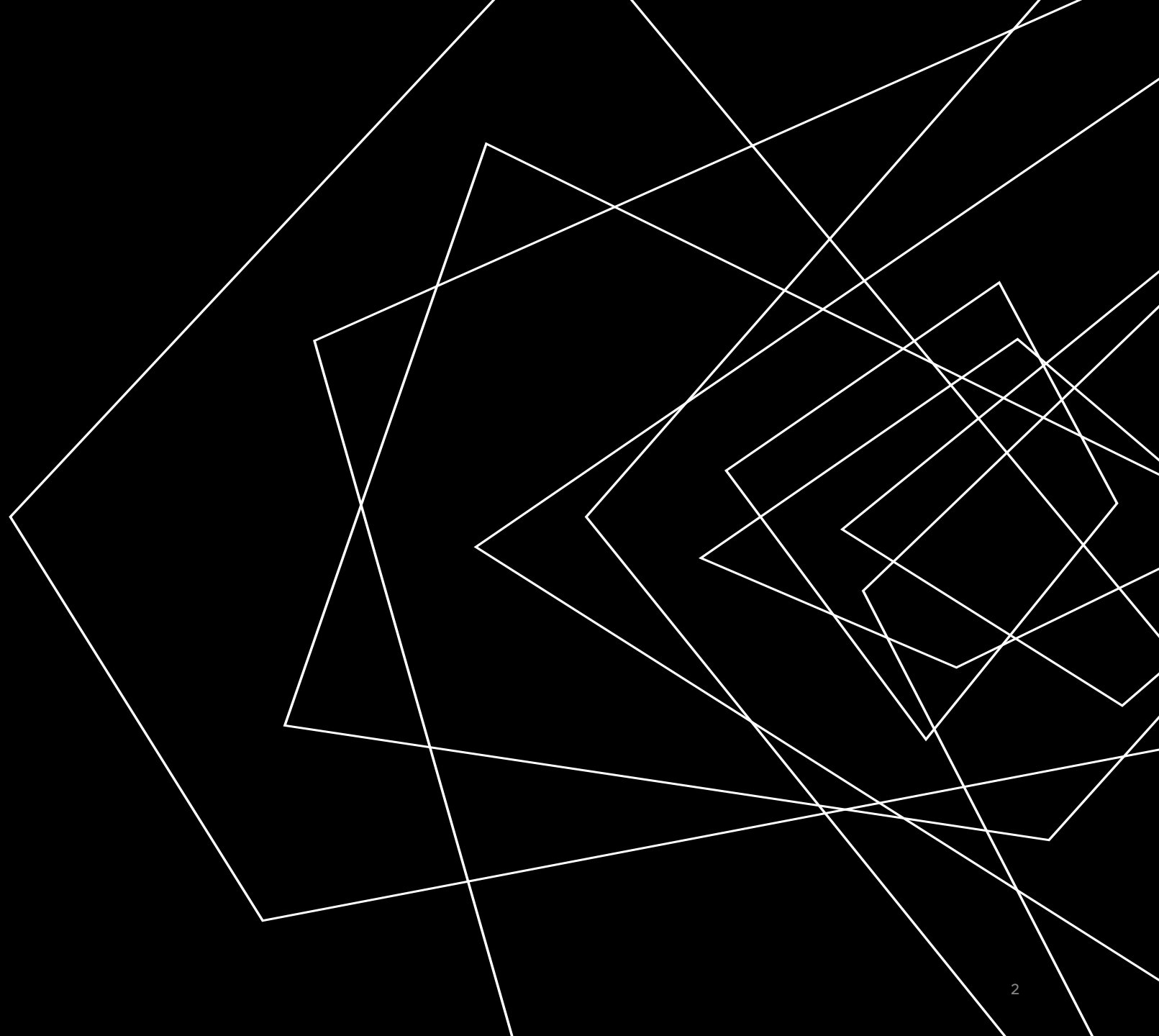
Introduction/Background

Dataset Curation/Benchmarking

Proposed Method

Benchmark Results

Analysis and Observations



INTRODUCTION

BACKGROUND

Inputs

"Five volunteers handing out food"

Output

Key Considerations:

- Can any recent work perform this task?
- After what point do the methods usually fail (2,3,4,5 people?)
- Does there exist any benchmark which compares all the methods performing this task?
- Is it an important problem in terms of usability?

DATASET CURATION

I. Curation

LLM Questions

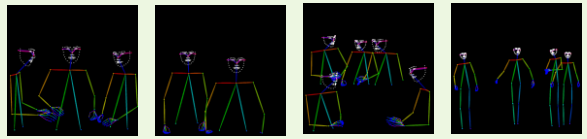
"Q1.: Provide 500 prompts placing 1-5 humans in the same frame?

Q2.: Provide 100 prompts placing 2-5 humans in the same frame performing different actions?"

LLM

Human Rectification

600 Pose Images



Generation +Selection

600 Prompts

100 Complex Prompts

"Five people at a picnic. One is holding a sandwich, four are raising drinks"

500 Simple Prompts

"Two hikers in a forest."

"Three chefs cooking a meal."

"Four friends in a coffee shop."

Prompt-Face Matching

Multi-Human Testbench

- 1800 Prompts
- 1800 Pose
- 5550 Faces

MLLM Questions

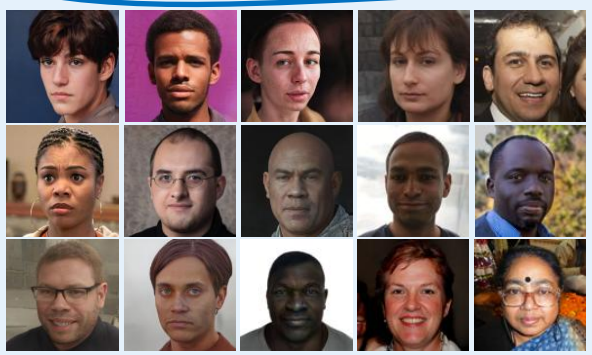
"Q: Estimate age, demography, gender?"

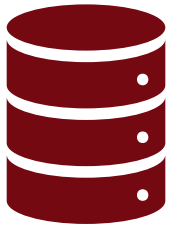
MLLM

Stratified Sampling

Open-Source Face Datasets (520K samples)

5500 Diverse Faces

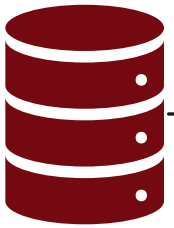




Multi-Human Testbench

- 1800 Prompts
- 1800 Pose
- 5550 Faces

II. Benchmarking



VLM

Multi-Human Testbench

- 1800 Prompts
- 1800 Pose
- 5550 Faces

Generated Images

Proposed Metrics

Measure face preservation
Hungarian ID similarity

Measure accurate face count
Detection+Accuracy

Measure prompt alignment
Human Preferance Score (HPSv2)

Measure simple actions
LMM-QA (Simple Actions)

Measure complex actions
LMM-QA (Complex Actions)

INITIAL OBSERVATIONS

Inputs

"Four hikers in the forest"

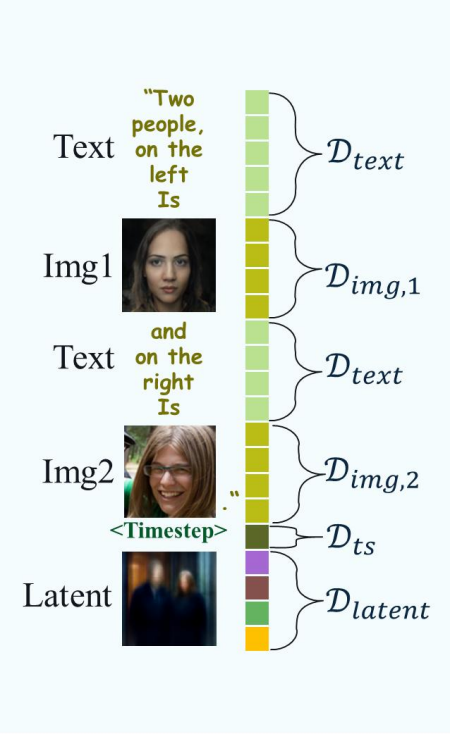
Output

SOTA methods fail due to:

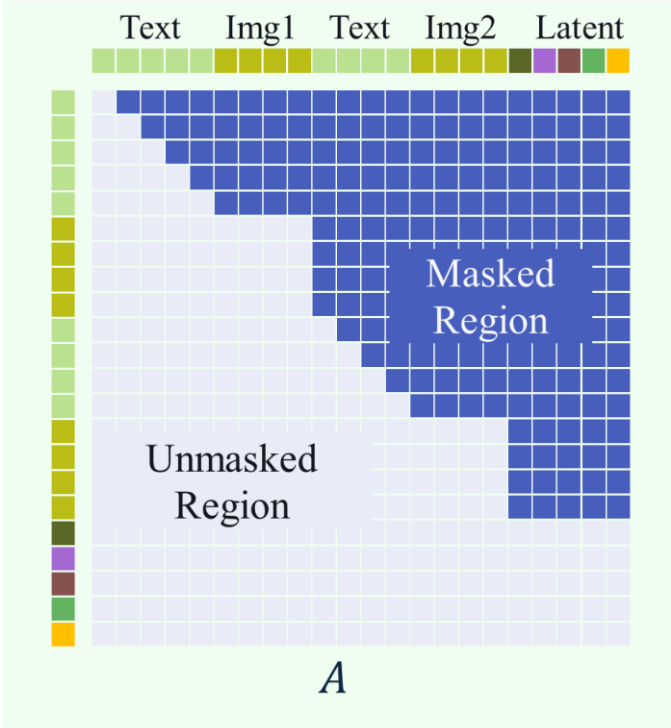
- Inaccurate number of generated people
- ID overlaps
- Text-described actions are not present

PROPOSED METHOD

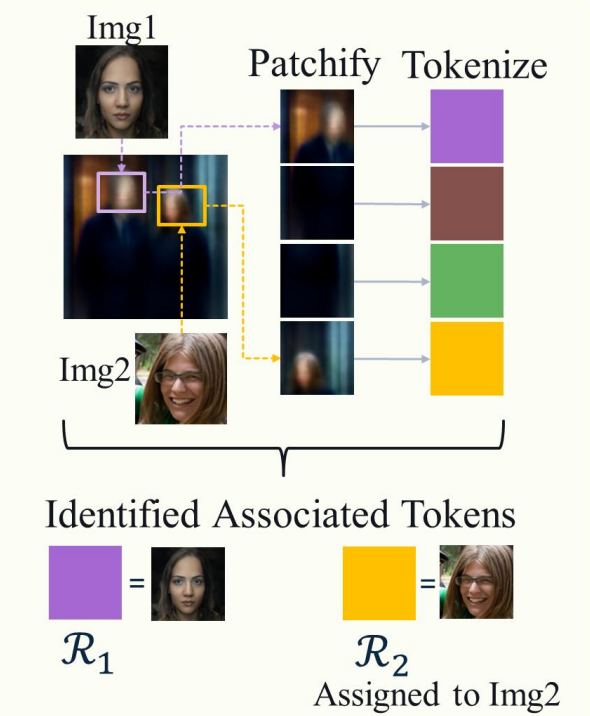
MULTIHUMAN OMNIGEN



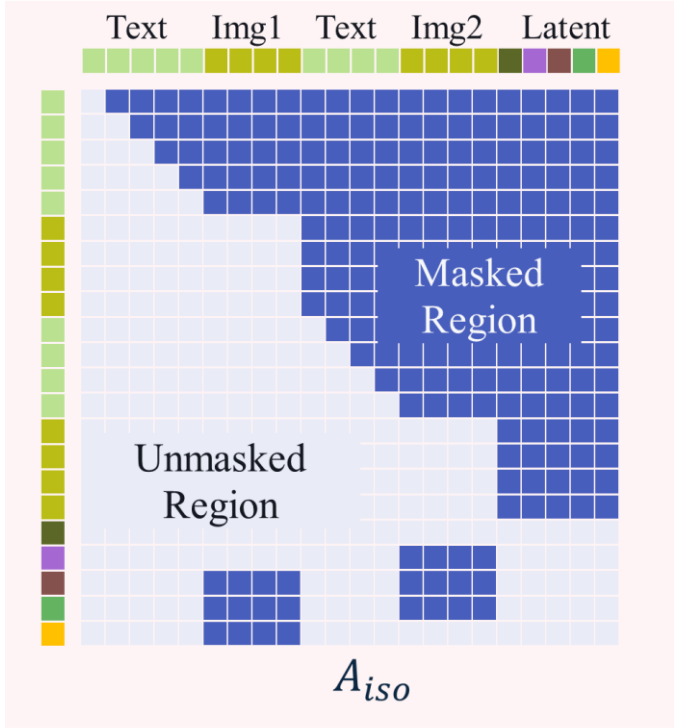
a. Token space



b. Attention Mask (Unified)



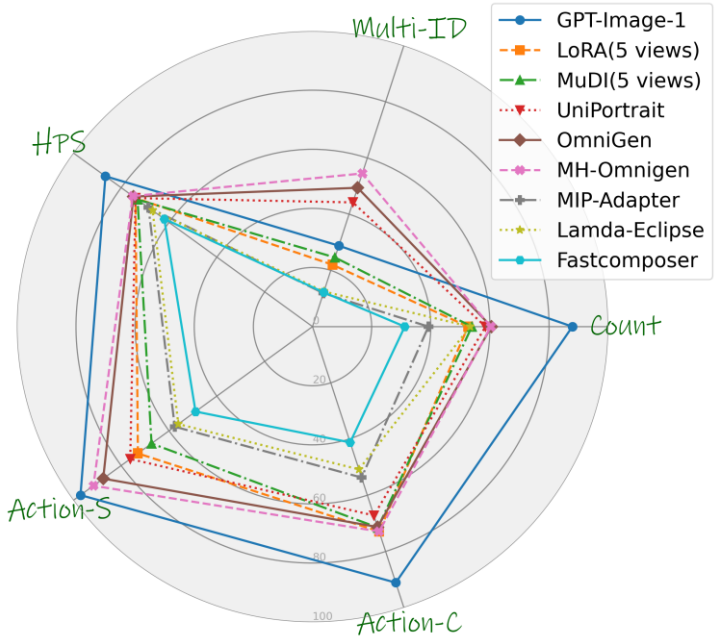
c. Regional Association

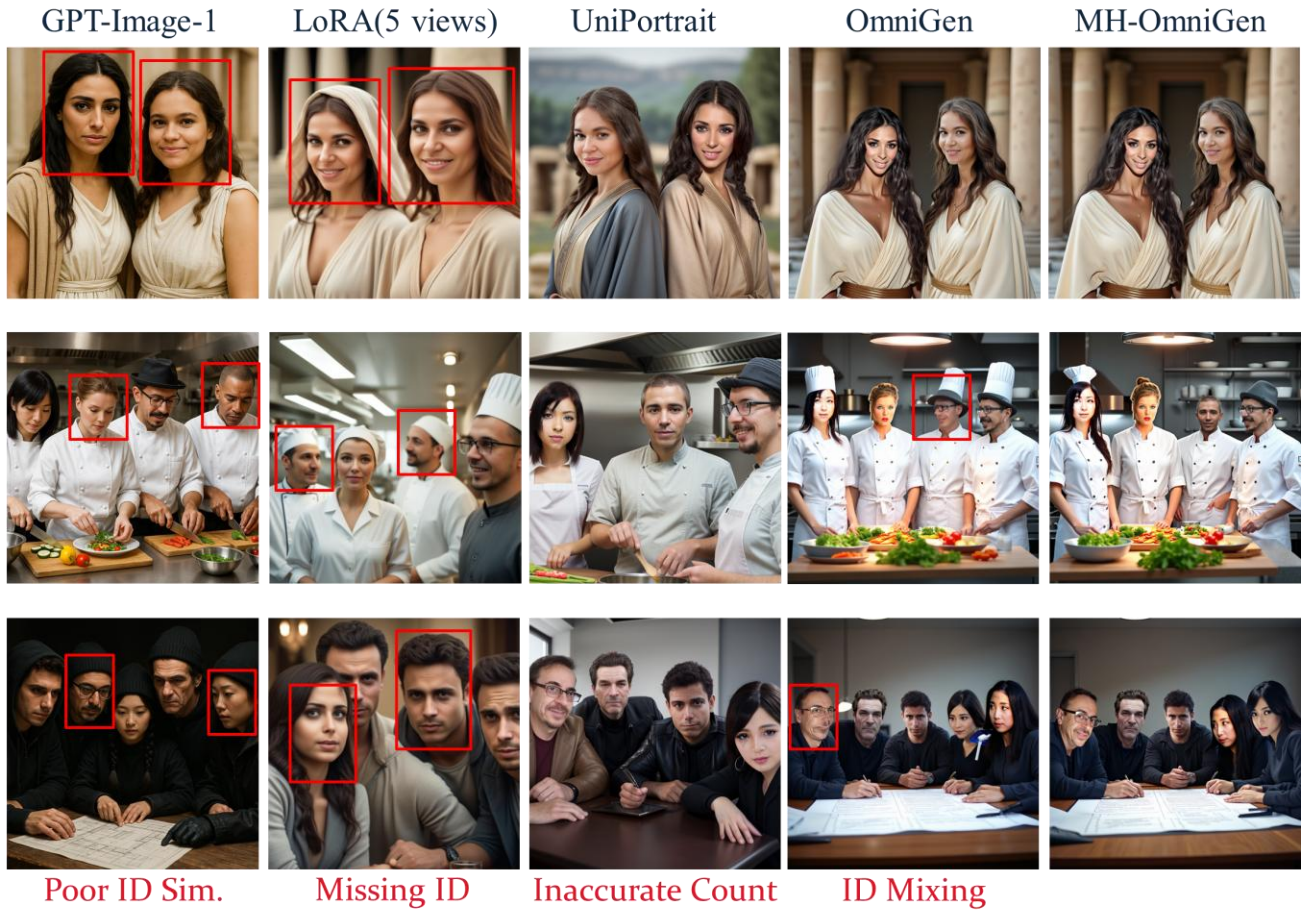


d. Unified Regional Isolation

RESULTS

REFERENCE-BASED MULTI-HUMAN GENERATION

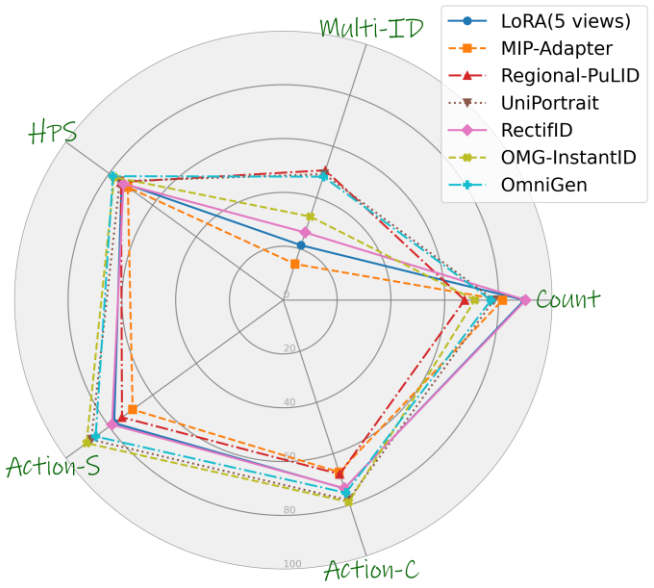


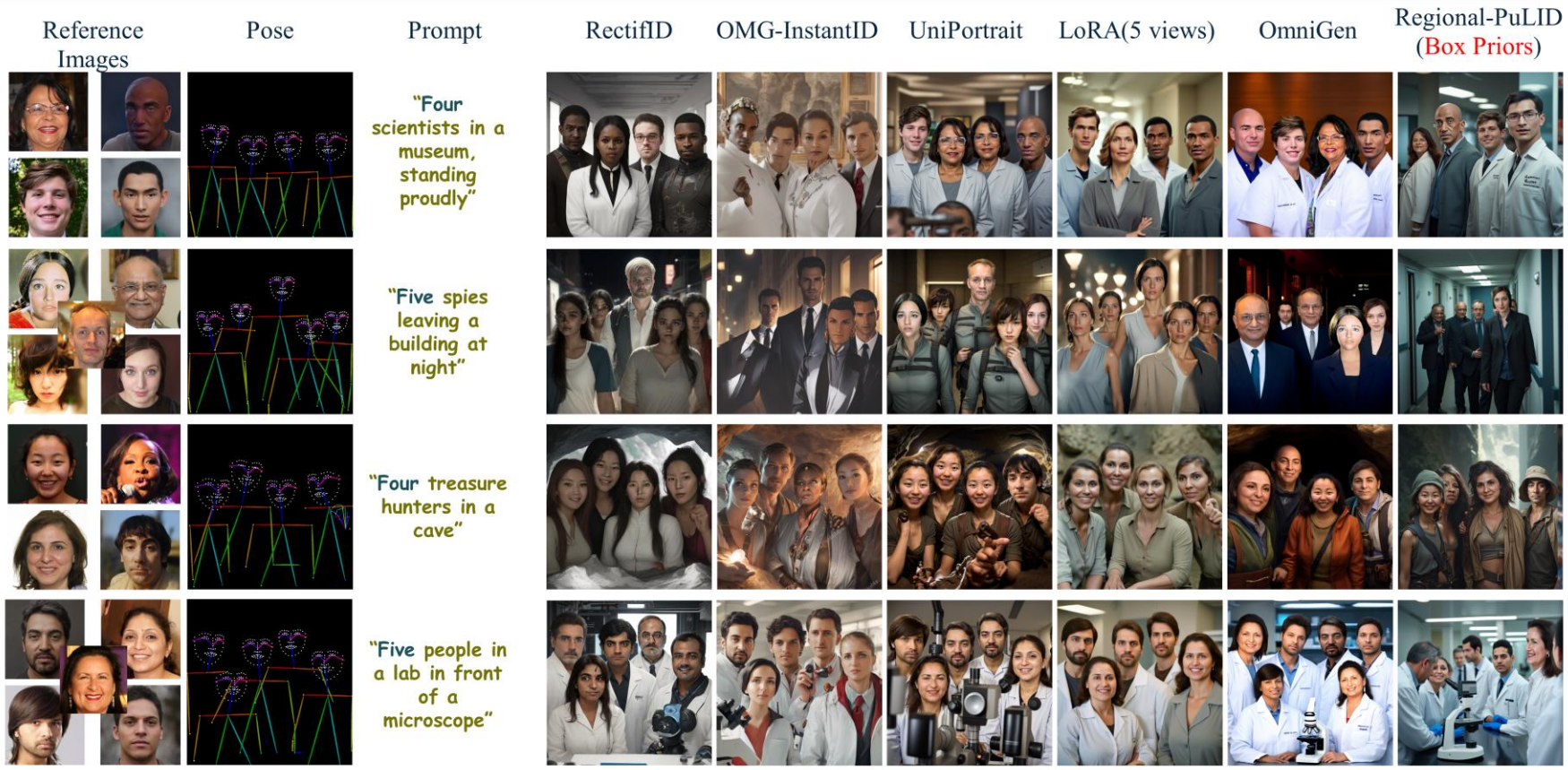


MULTIHUMAN-OMNIGEN BENEFIT

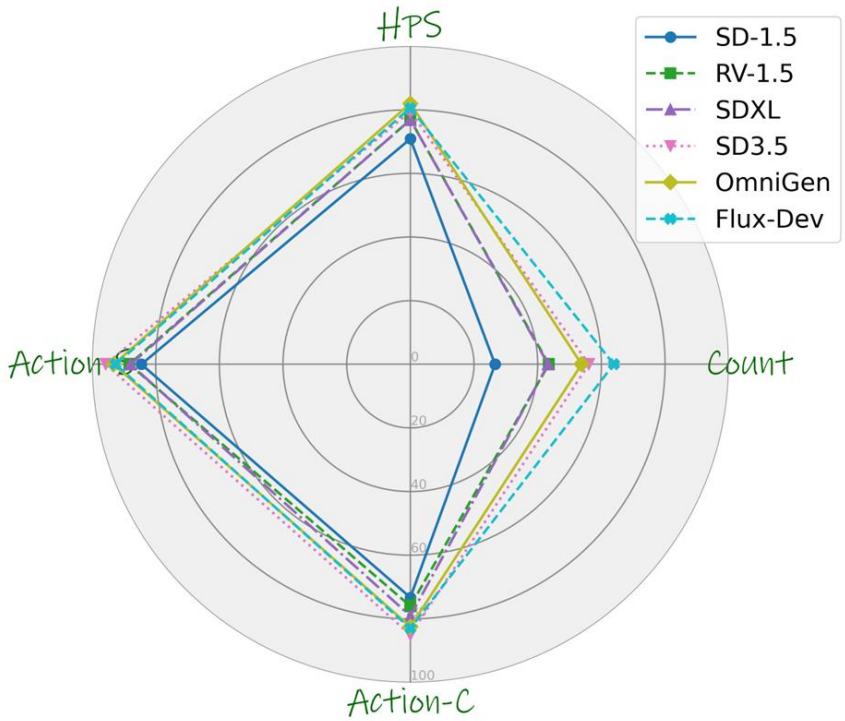
	Reference Images	OmniGen	MH-OmniGen
"Three chefs cooking a meal"			
"Three firefighters putting out a raging fire"			
"Four treasure hunters finding gold in a cave"			

REF. BASED MULTI-HUMAN GENERATION WITH SPATIAL PRIOR





TEXT-TO-IMAGE MULTI-HUMAN GENERATION



Prompt	RV1.5	SDXL	SD3.5	OmniGen	Flux
"Five people during Diwali with sparklers and lanterns, fireworks in the sky"					
"Five spies side by side escaping danger; running, explosions in the background;"					
"Four friends side by side in a jungle; dense vegetation; adventure; sunlight."					
"Four students in a library; books, laptops."					

CONCLUSION