

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

An abstract geometric design featuring two thin, dark grey lines that intersect on a light grey background. One line is oriented diagonally from the top-left towards the bottom-right, while the other is oriented from the top-right towards the bottom-left. The intersection point is located in the upper-left quadrant of the frame.

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?

An abstract geometric design featuring two thin, dark lines that intersect on a light gray background. One line is oriented diagonally from the top-left towards the bottom-right, while the other is oriented from the top-right towards the bottom-left. The intersection point is located to the left of the text.

DATASET CURATION

I. Curation

LLM Questions

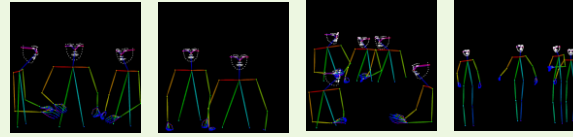
"Q1.: Provide **500 prompts** placing 1-5 humans in the same frame?"

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



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**"

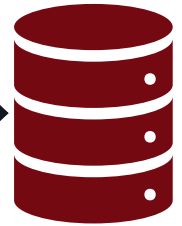
500 Simple Prompts

"**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?"

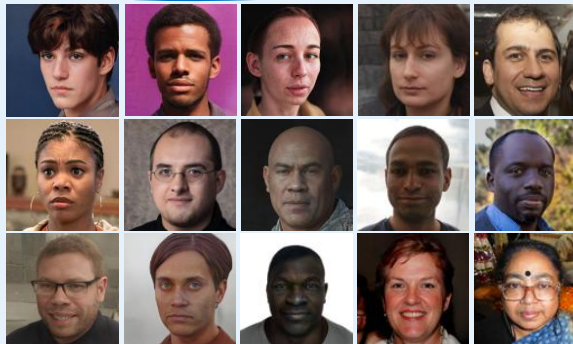


Open-Source
Face Datasets
(520K samples)



Stratified
Sampling

5500 Diverse Faces

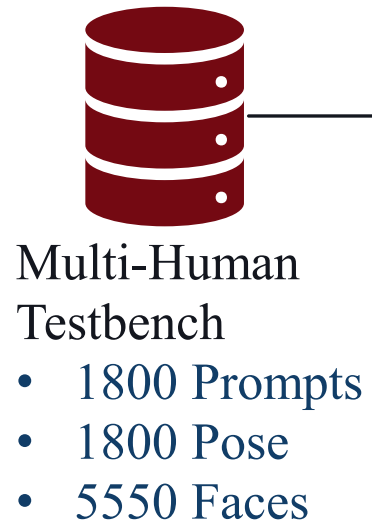




Multi-Human Testbench

- 1800 Prompts
- 1800 Pose
- 5550 Faces

II. Benchmarking



Proposed Metrics

- Measure face preservation
Hungarian ID similarity
- Measure accurate face count
Detection+Accuracy
- Measure prompt alignment
Human Preference 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:

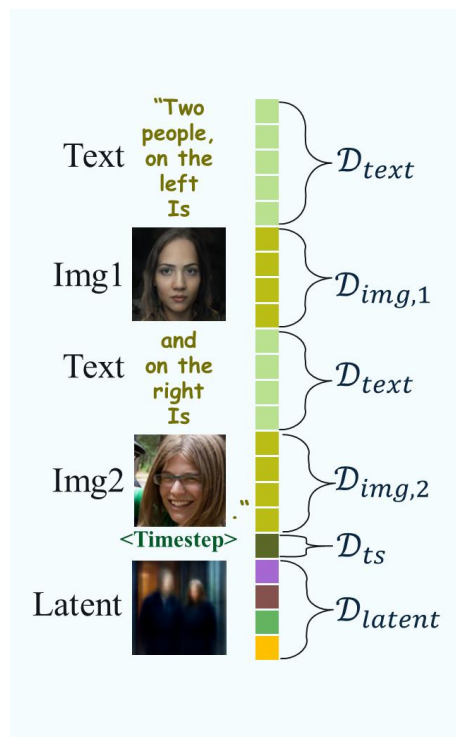
- Inaccurate number of generated people
- ID overlaps
- Text-described actions are not present



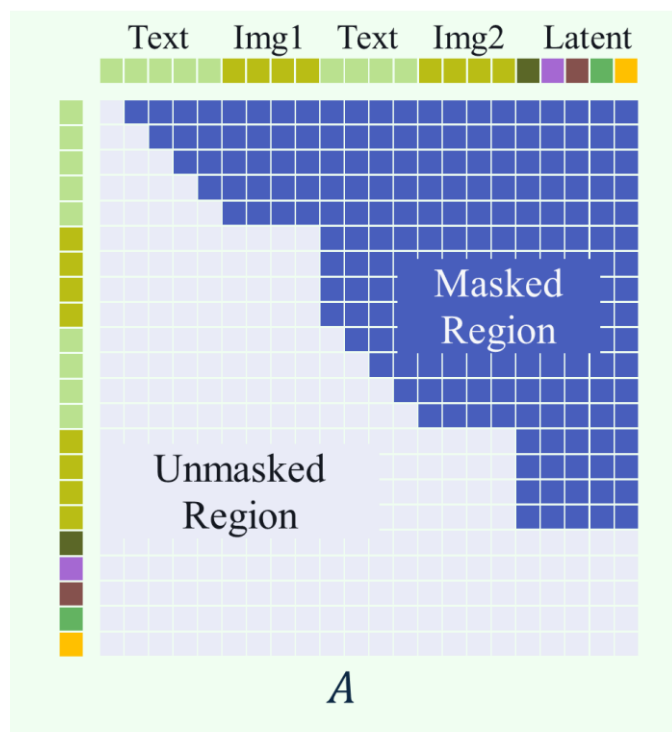
PROPOSED
METHOD

The image features a minimalist design on a light gray background. Two thin, dark gray lines intersect: one line runs diagonally from the top-left towards the bottom-right, and the other runs from the top-right towards the bottom-left. To the right of this intersection, the words 'PROPOSED' and 'METHOD' are stacked vertically in a bold, black, sans-serif typeface.

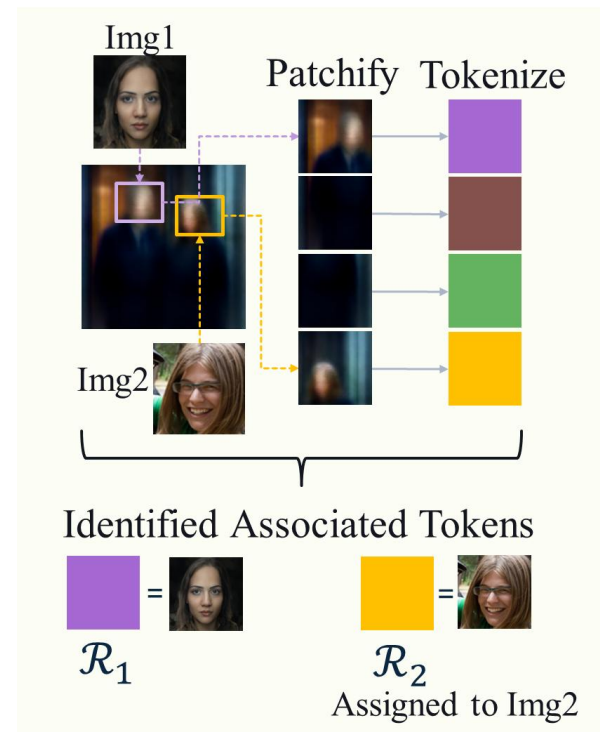
MULTIHUMAN OMNIGEN



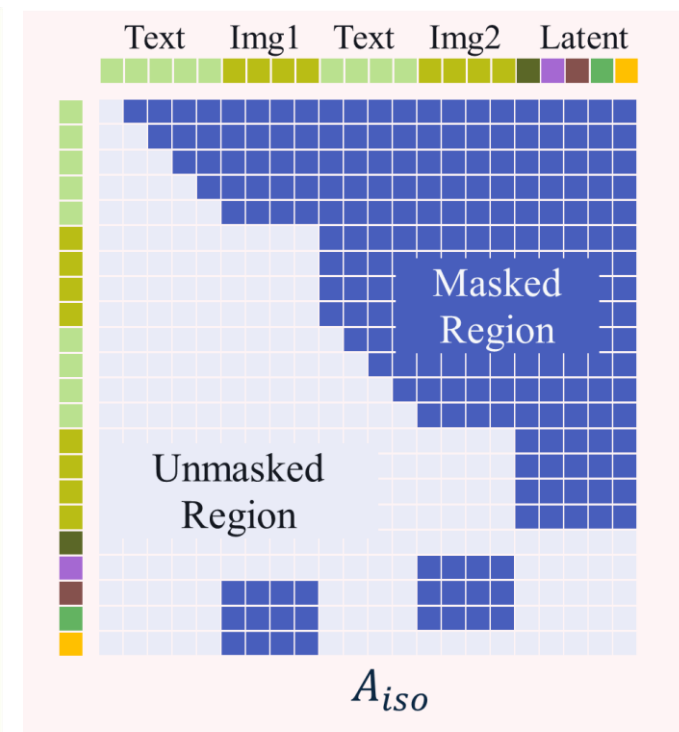
a. Token space



b. Attention Mask (Unified)



c. Regional Association

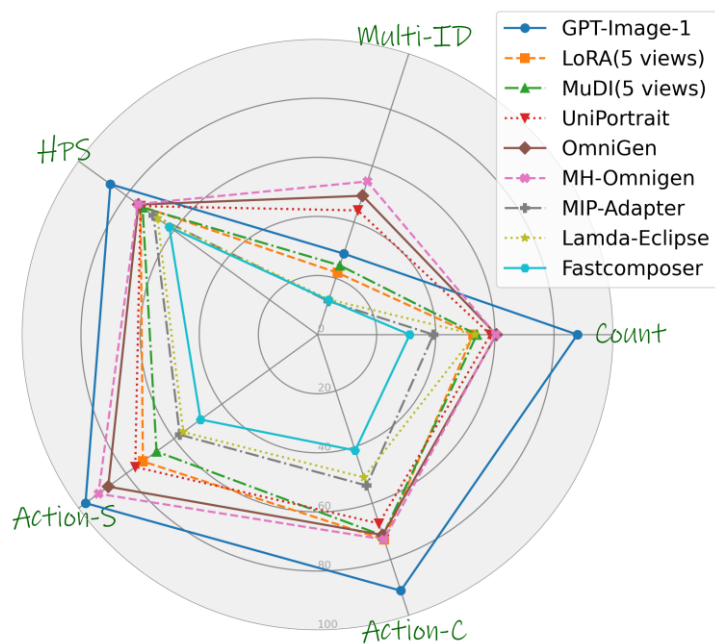


d. Unified Regional Isolation

An abstract graphic featuring two thin, dark grey lines that intersect on a light grey background. One line is oriented diagonally from the top-left towards the bottom-right, while the other is oriented from the top-right towards the bottom-left. The word "RESULTS" is positioned to the right of the intersection point, rendered in a bold, black, sans-serif typeface.

RESULTS

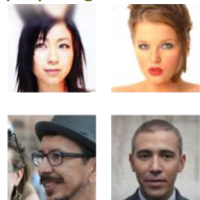
REFERENCE-BASED MULTI-HUMAN GENERATION



Inputs
"Two people in ancient Greece wearing robes"



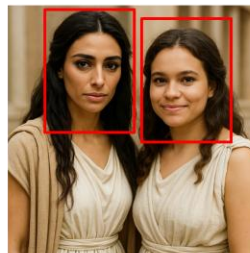
"Four chefs in a kitchen preparing a meal"



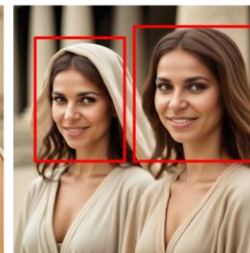
"Five thieves planning a heist"



GPT-Image-1



LoRA(5 views)



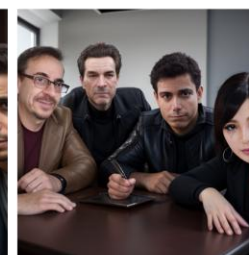
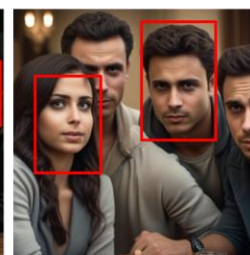
UniPortrait



OmniGen



MH-Omnigen



Poor ID Sim.

Missing ID

Inaccurate Count

ID Mixing

MULTIHUMAN-OMNIGEN BENEFIT

Reference Images

"Three
chefs
cooking a
meal"



OmniGen



MH-OmniGen



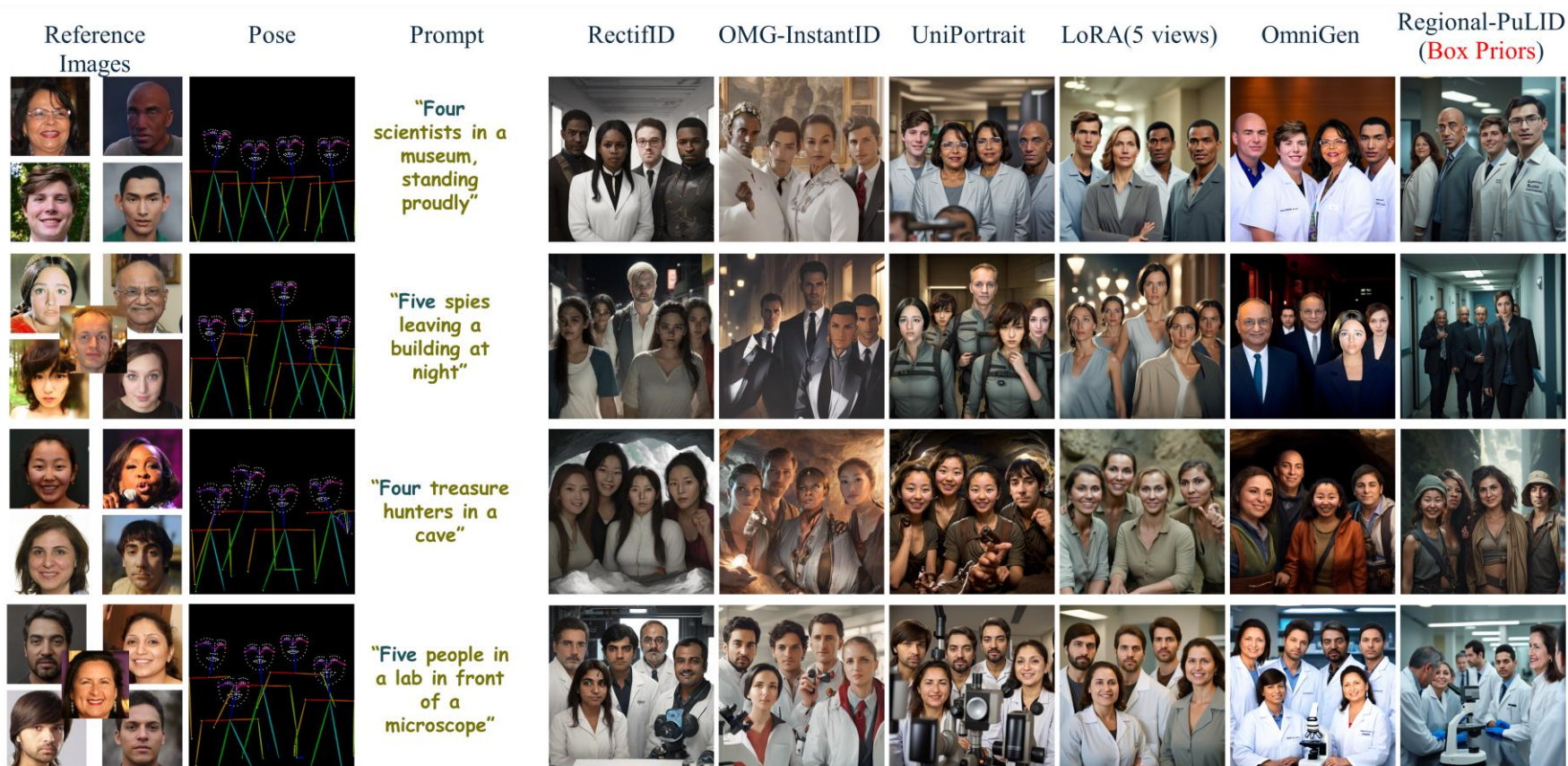
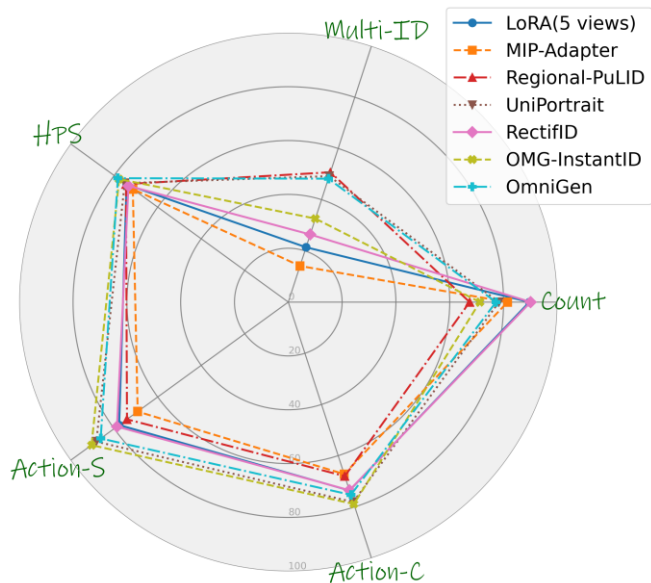
"Three
firefighters
putting out
a raging
fire"



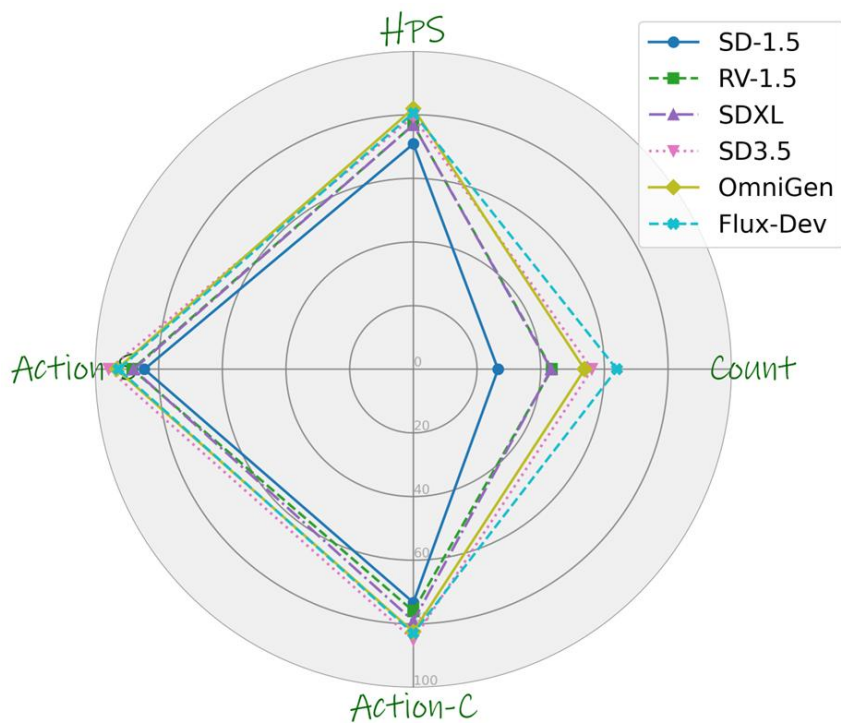
"Four
treasure
hunters
finding gold
in a cave"



REF. BASED MULTI-HUMAN GENERATION WITH SPATIAL PRIOR



TEXT-TO-IMAGE MULTI-HUMAN GENERATION



Prompt

"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."

RV1.5



SDXL



SD3.5



OmniGen



Flux



An abstract graphic featuring two thin, dark grey lines that intersect on a light grey background. One line is oriented diagonally from the top-left towards the bottom-right, while the other is more horizontal, sloping slightly downwards from left to right. The intersection point is located in the upper-left quadrant of the image. To the right of this intersection, the word "CONCLUSION" is written in a bold, black, sans-serif font.

CONCLUSION