Qualcomm

Ar2Can: An Architect and an Artist Leveraging a Canvas for Multi-Human Generation

Shubhankar Borse, Phuc Pham, Farzad Farhadzadeh, Seokeon Choi, Phong Ha Nguyen, Anh Tuan Tran, Sungrack Yun, Munawar Hayat, Fatih Porikli

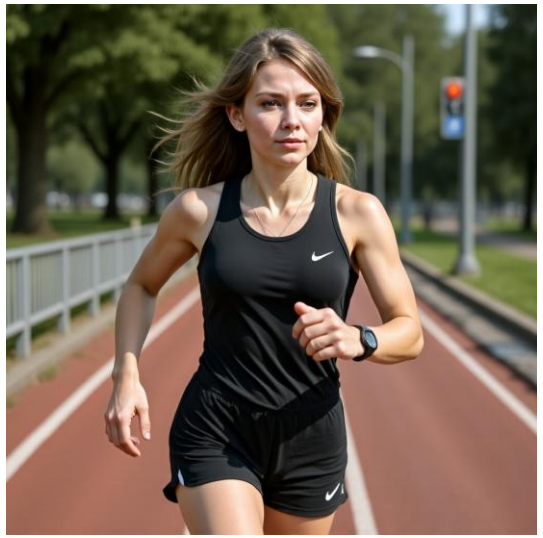
https://qualcomm-ai-research.github.io/ar2can/

Presenter: Shubhankar Borse
Staff AI Reseacher
Qualcomm AI Reserach

@qualcomm

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patents are licensed by Qualcomm Incorporated.

MOTIVATION



We aim to solve the problem posed by MultiHuman-Testbench paper.* Given input references of multiple – people, we want to generate personalized images of them in the same scene which are

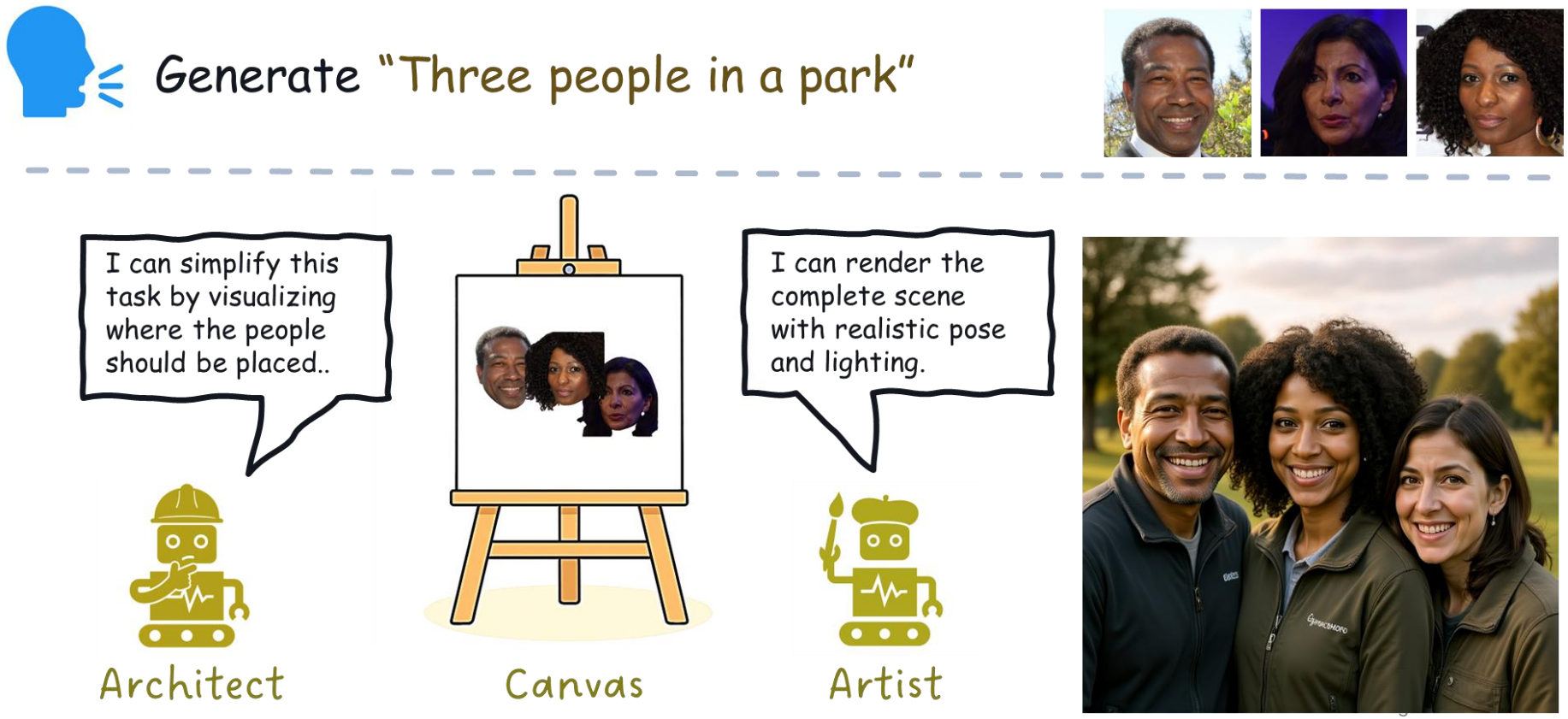
- Photorealistic
- Identity Preserving
- Prompt-Following

*Source: https://arxiv.org/abs/2506.20879

SUMMARY

We developed a two-stage approach for generating a group scene.

- The Architect generates a spatial layout for our scene to simplify the task.
- The Artist uses this spatial layout as a reference to perform generation.
- We develop a framework to finetune our architect and artist using RL-based GRPO.

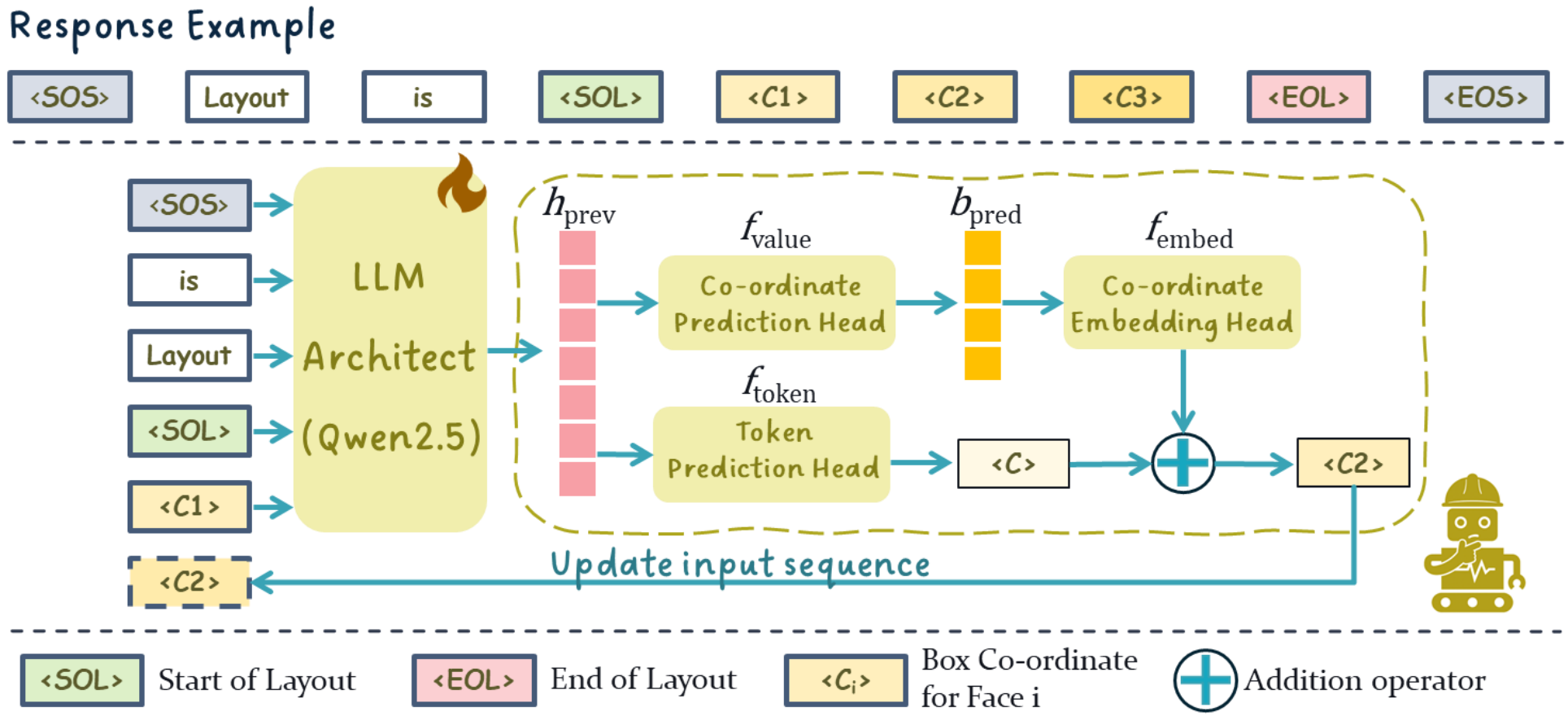


Proposed Method

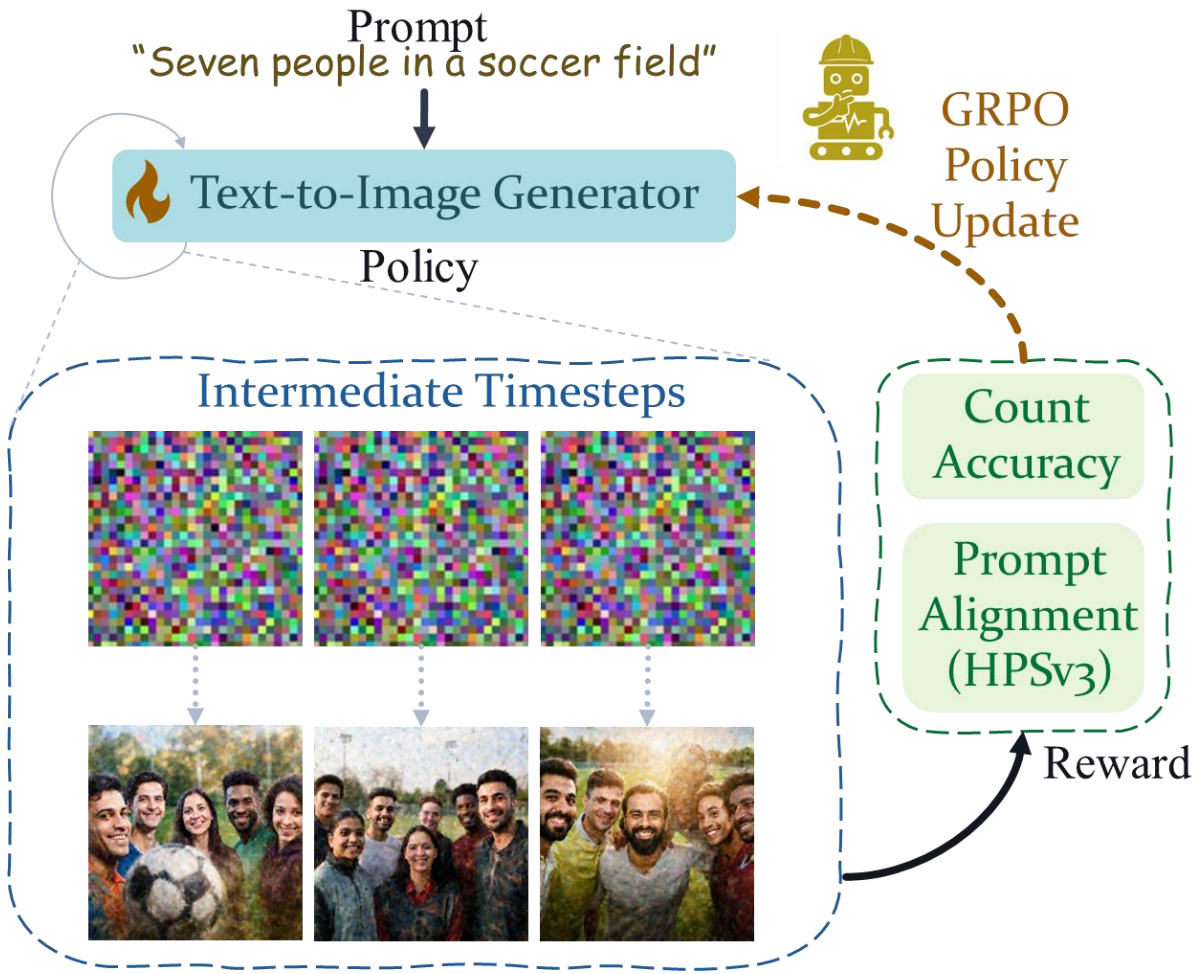
ARCHITECT-A

We build an autoregressive language model that predicts spatial layouts from text prompts.

- We fine-tune Qwen2.5-0.5B with supervised learning on layout tokens and coordinate regression, ensuring accurate placement in multi-human scenes.

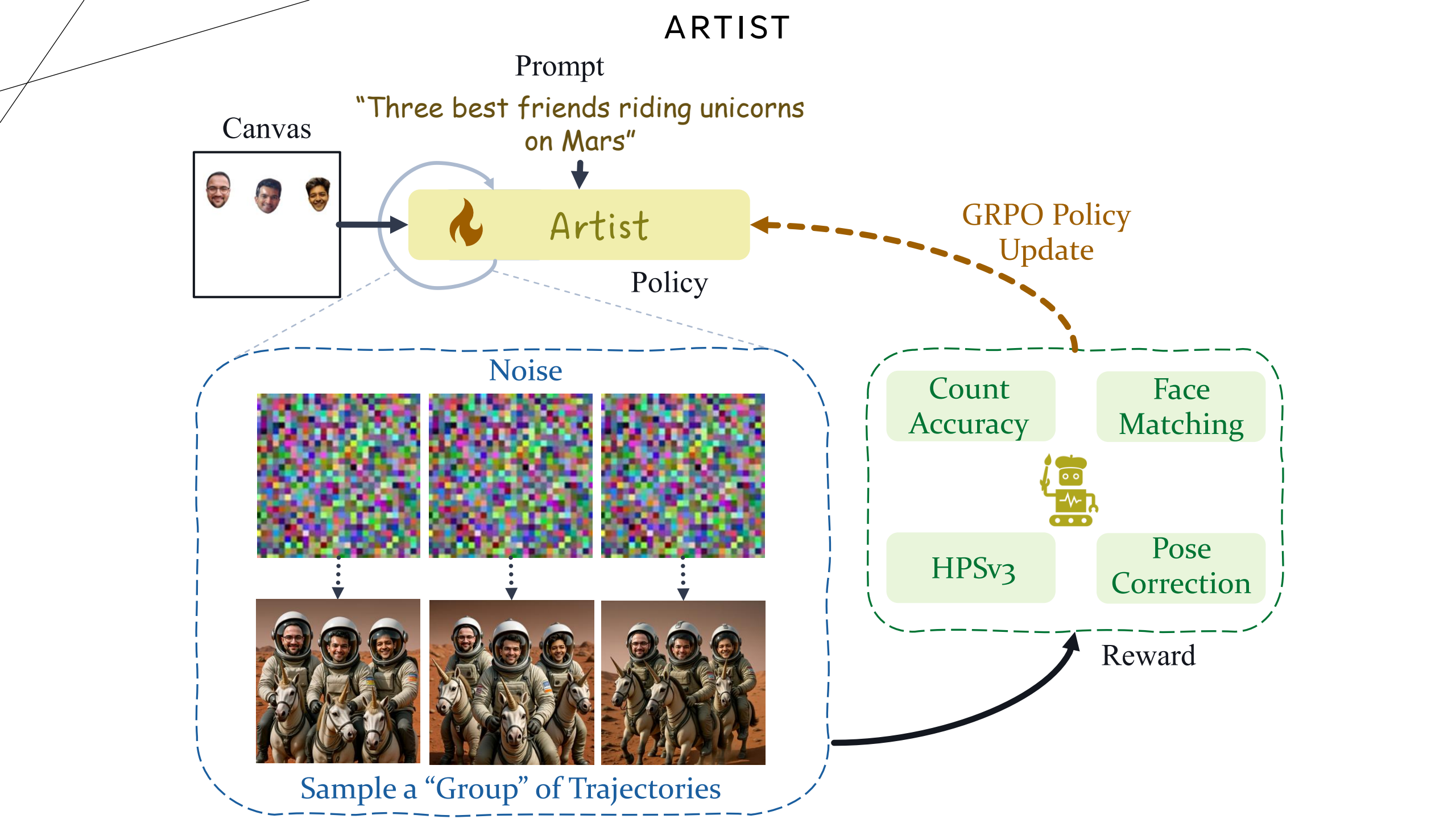


ARCHITECT-B



We build another architect variant using fast T2I diffusion model.

- We fine-tune Flux-Schnell with reinforcement learning to generate the correct number of people in a proper layout based on the prompt.
- We extract the positions of every face from this fine-tuned model.

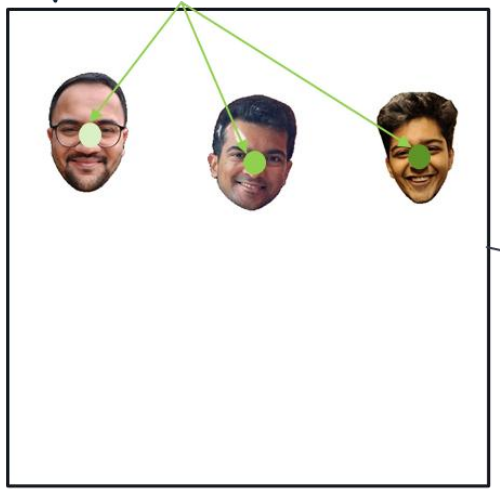

ARTIST
Prompt
"Three best friends riding unicorns on Mars"
Canvas
Artist
Policy
GRPO Policy Update
Noise
Count Accuracy
Face Matching
HPSv3
Pose Correction
Reward
Sample a "Group" of Trajectories

ARTIST

Face Matching

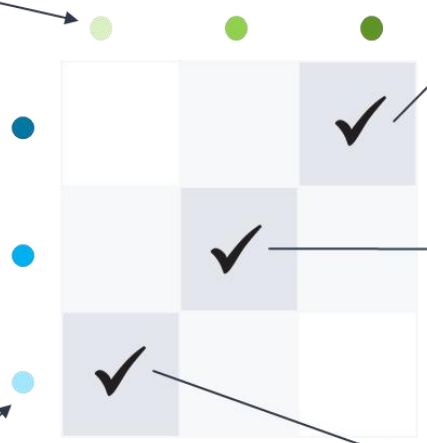
Face Matching

Input Centroids



Generated Centroids

Hungarian Matching



C

Cost-Matrix

$I_{\text{ref},1}$ 

$I_{\text{ref},2}$ 

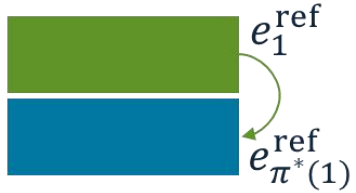
$I_{\text{ref},3}$ 

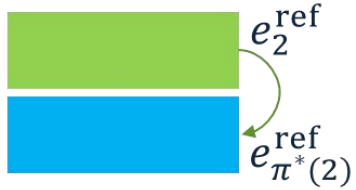
Arcface

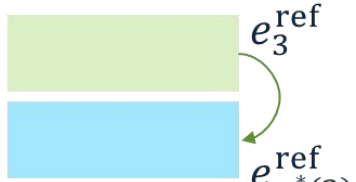
Arcface

Arcface

Embeddings



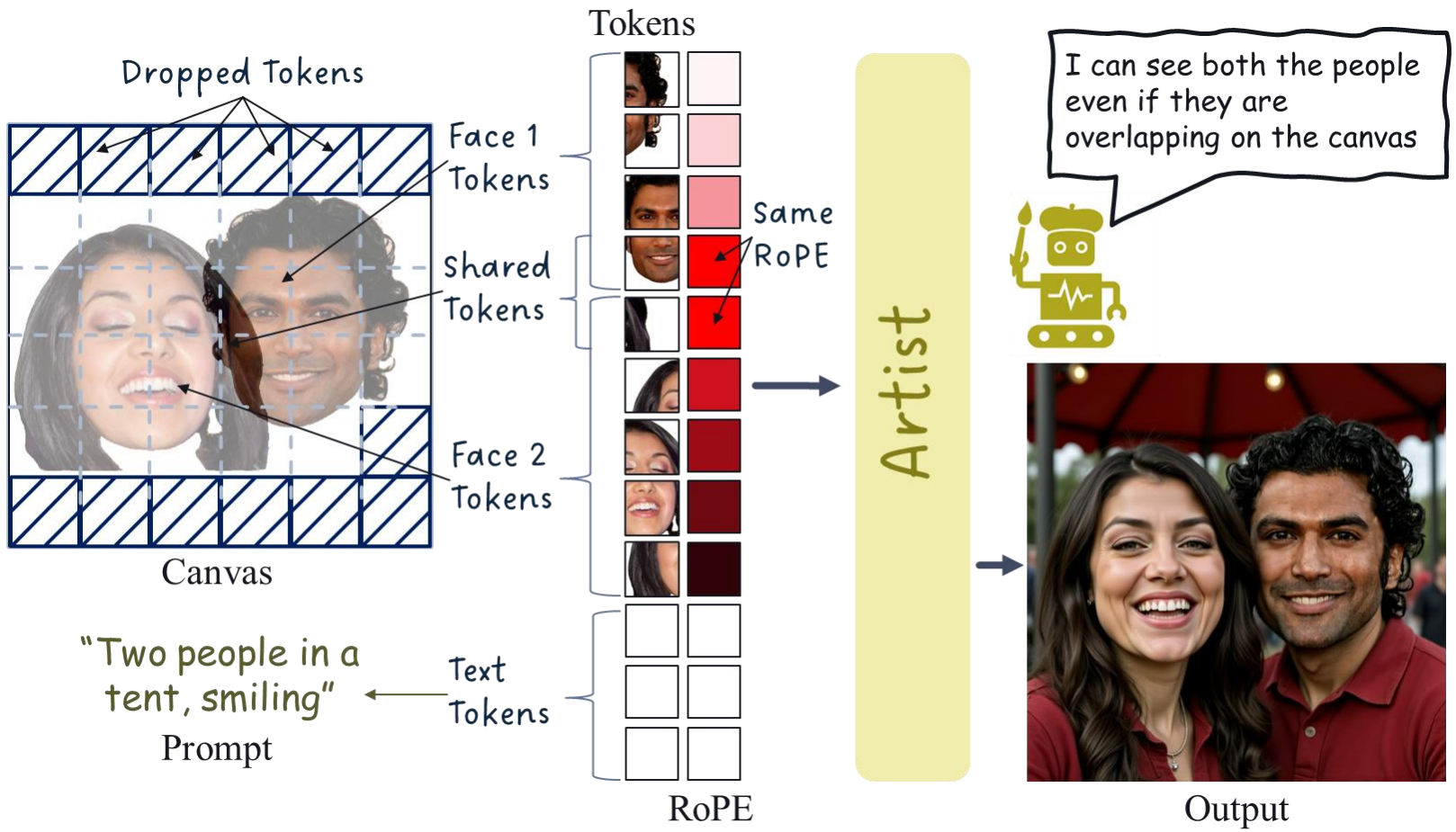




$$r_{\text{face}} = \frac{1}{3}\sum_{i=1}^{3} \cos(e_i^{\text{ref}}, e_{\pi^*(i)}^{\text{ref}})$$

TOKEN COMPRESSION/OVERLAPPING

- We accelerate inference by dropping empty canvas regions and sharing tokens for overlapping areas, cutting computation nearly in half.
- Shared positional encodings allow natural occlusion handling and depth ordering, enabling realistic multi-human layouts without extra cost.

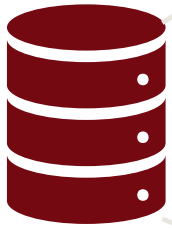


Experiments and Results

TEST DATASET

Multi-Human Testbench

- 1800 samples
- 1-5 People



Multi-ID Similarity: Face Detection + Hungarian Matching + ArcFace Similarity

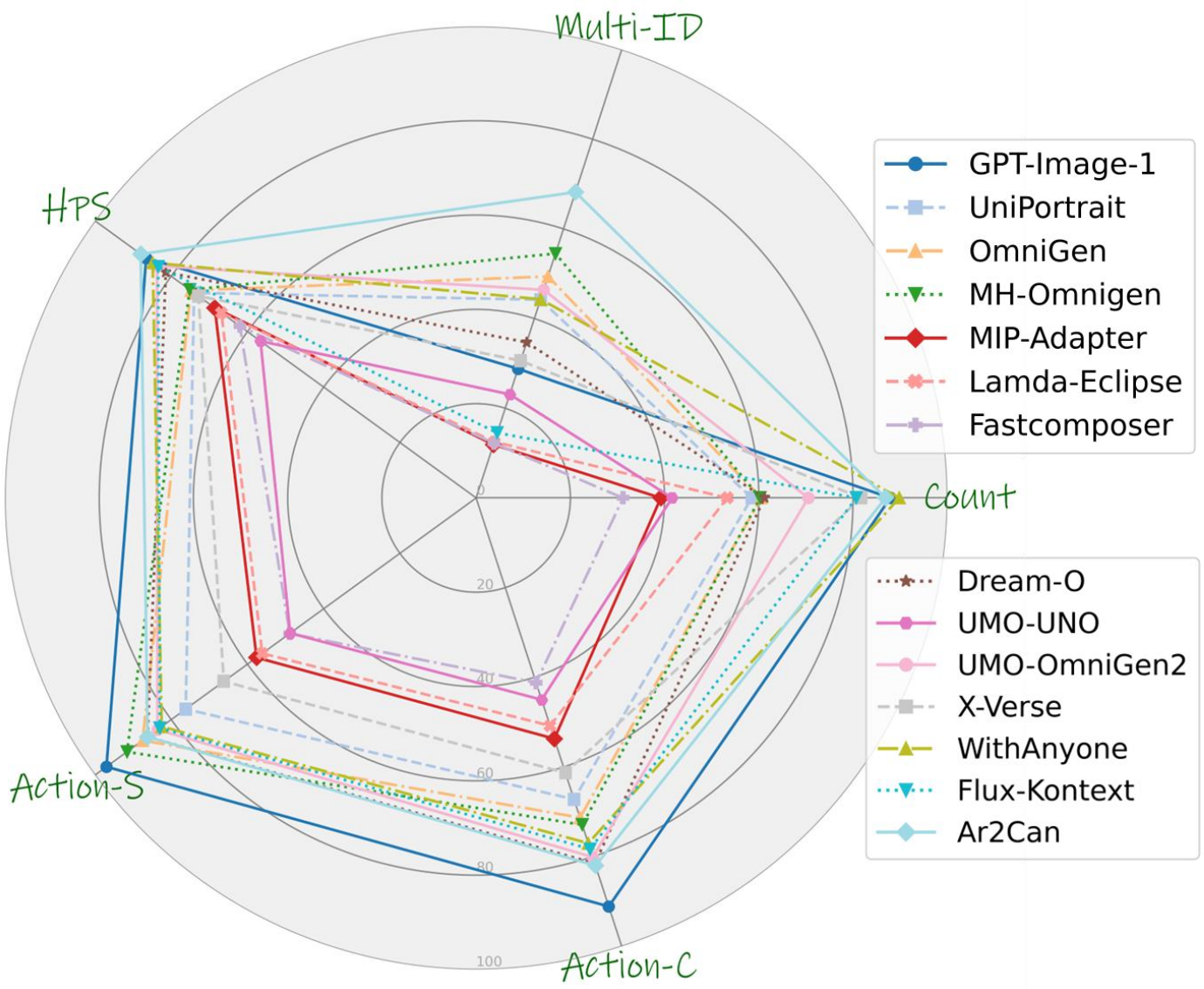
Person Count Accuracy: Percentage of images with correct number of faces

Hpsv2: Human Preference Score based VLM model. High HPS → More prompt overlap + More Realistic

Simple Action Score: MLLM-based Action verification score (0-100) for simple prompts

Complex Action Score: MLLM-based Action verification score (0-100) for complex prompts

QUANTITATIVE RESULTS



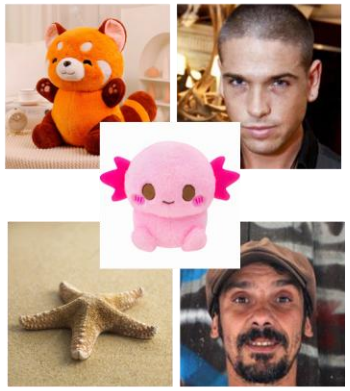
Observations:

- Ar2Can achieves **best spread** across all metrics.
- Ar2Can retains identity while keeping quality and prompt alignment high.
- Proprietary methods (GPT, Gemini) perform well on Action scores

QUALITATIVE RESULTS

Scoresheet: All IDs Retained: ✓ Prompt Action Aligned: ✓ IDs not retained/Prompt not aligned: ✗

QUALITATIVE RESULTS

Prompt	Inputs	Output
"Two men sitting on a log on the beach. Two plushies are on either side of them. The person on the left wears a green hawaiian shirt and barefoot. The person on the right wears sunglasses and is laughing. The plushie on the right is wearing a black hat. A starfish is on the sand."		

Observations:

- Even though we train only for human faces, Ar2Can can successfully perform Multi-Object + Multi-Person Generation.

Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated.
Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patents are licensed by Qualcomm Incorporated.

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog

Multi-Huma

n, Farzad Farhadzadeh,
en, Anh Tuan Tran, Sungrack
rikli