NEURAL INFORMATION PROCESSING SYSTEMS

MUSLIMS IN MACHINE LEARNING

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

THE ALIGARH MUSLIM UNIVERSITY ALIGARH

Northeastern University

From Rules to Pixels

A Decoupled Framework for Segmenting Human-Centric Rule Violations

Mohd Hozaifa Khan, Harsh Awasthi, Pragati Jain, Mohammad Ammar

Problem Formulation

Objective:

Translate long-form textual appearance rules into accurate, pixel-level outputs that indicate which regions in an image do not satisfy the specified policy.

Motivation & Challenges:

- Many environments rely on written appearance guidelines that are contextual and culturally influenced
- Correct interpretation requires identifying specific body parts and checking if they are exposed or covered
- Current vision–language models only detect objects and lack explicit rule-logic execution
- Without this reasoning step, outputs become incorrect, non-auditable, and unreliable for deployment in human-centered settings

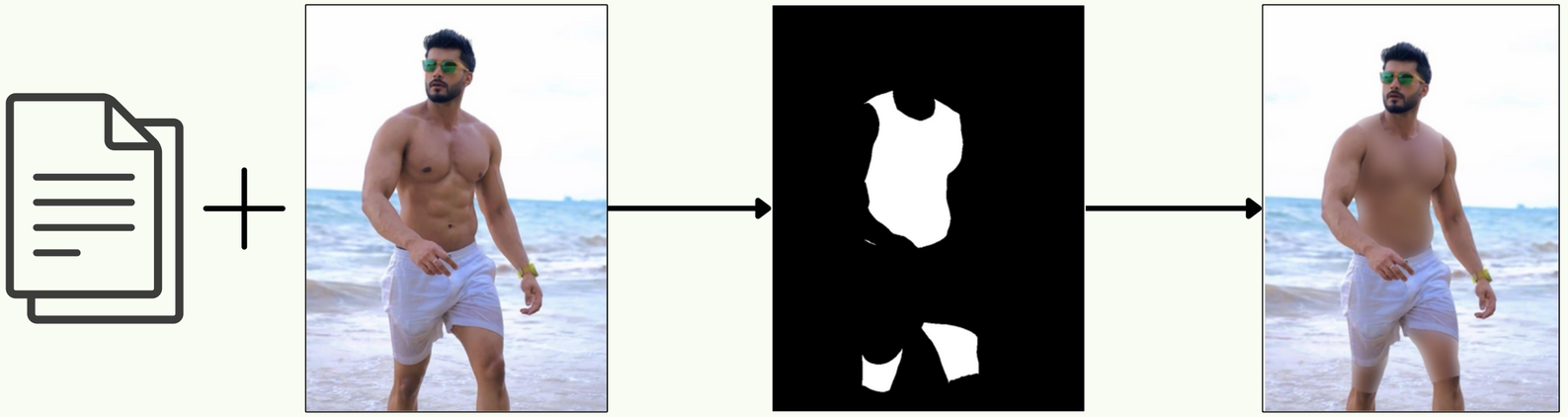


Image + textual appearance rule

Highlight inconsistent regions

Prior Works and Gap

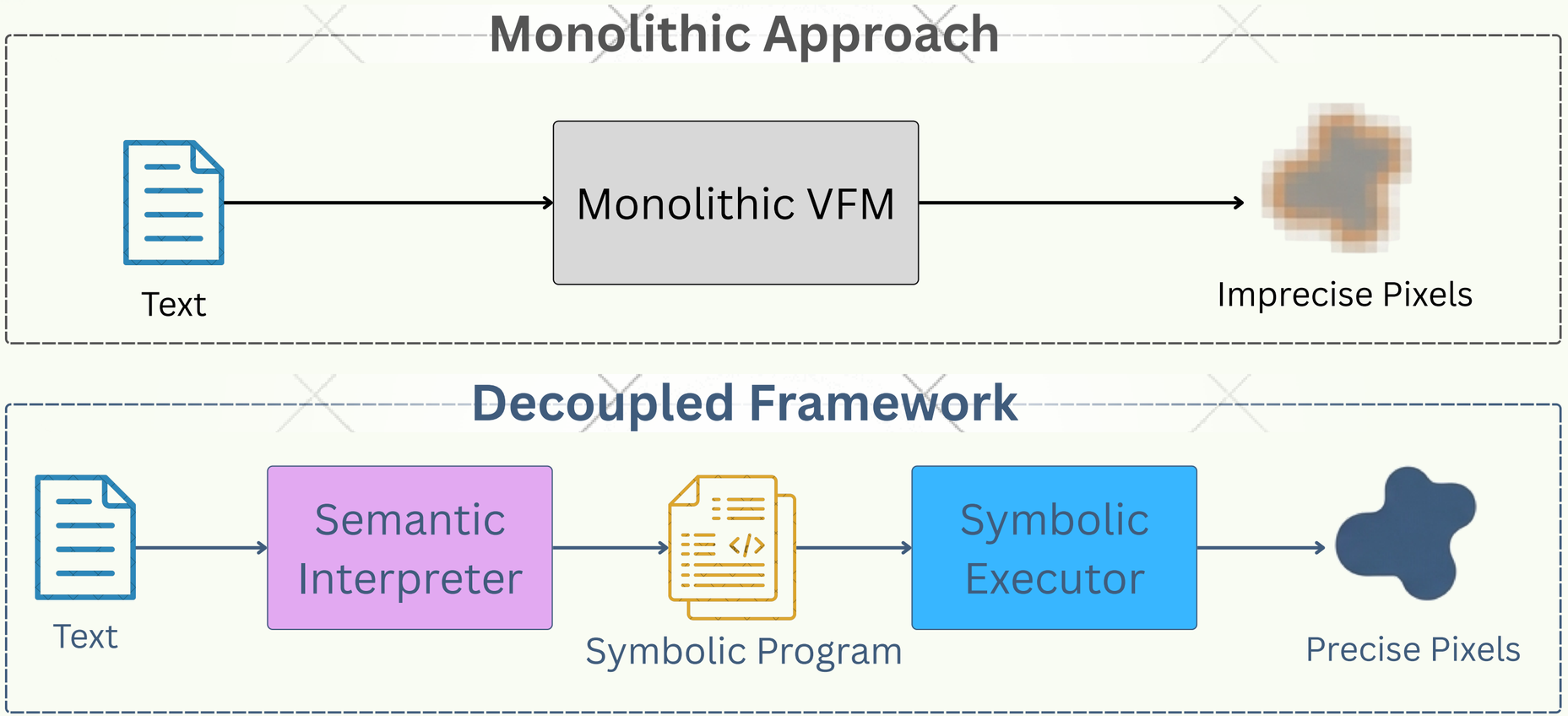
Limitations of Existing Approaches

- Grounding models focus on object matching, not rule interpretation
- VLMs cannot parse compositional logic in rules
- (e.g., “below the knee but above the ankle”)
- Their outputs are opaque, with no explanation of why a region is highlighted
- No benchmark evaluates human-centric rule grounding at pixel-level

How LaGPS Advances the Field

- Converts free-form rules into symbolic programs with explicit logical structure
- Executes program reasoning over visual primitives (body-parts, skin regions, etc.)
- Generates accurate and auditable pixel-level results
- Handles conditional, context-dependent semantics that prior VLMs cannot

Monolith vs Decoupled



Semantic - Symbolic Gap

	Logic	Vision
LLMs		
VFMs		

Key Idea

Interpretation ≠ **Execution**

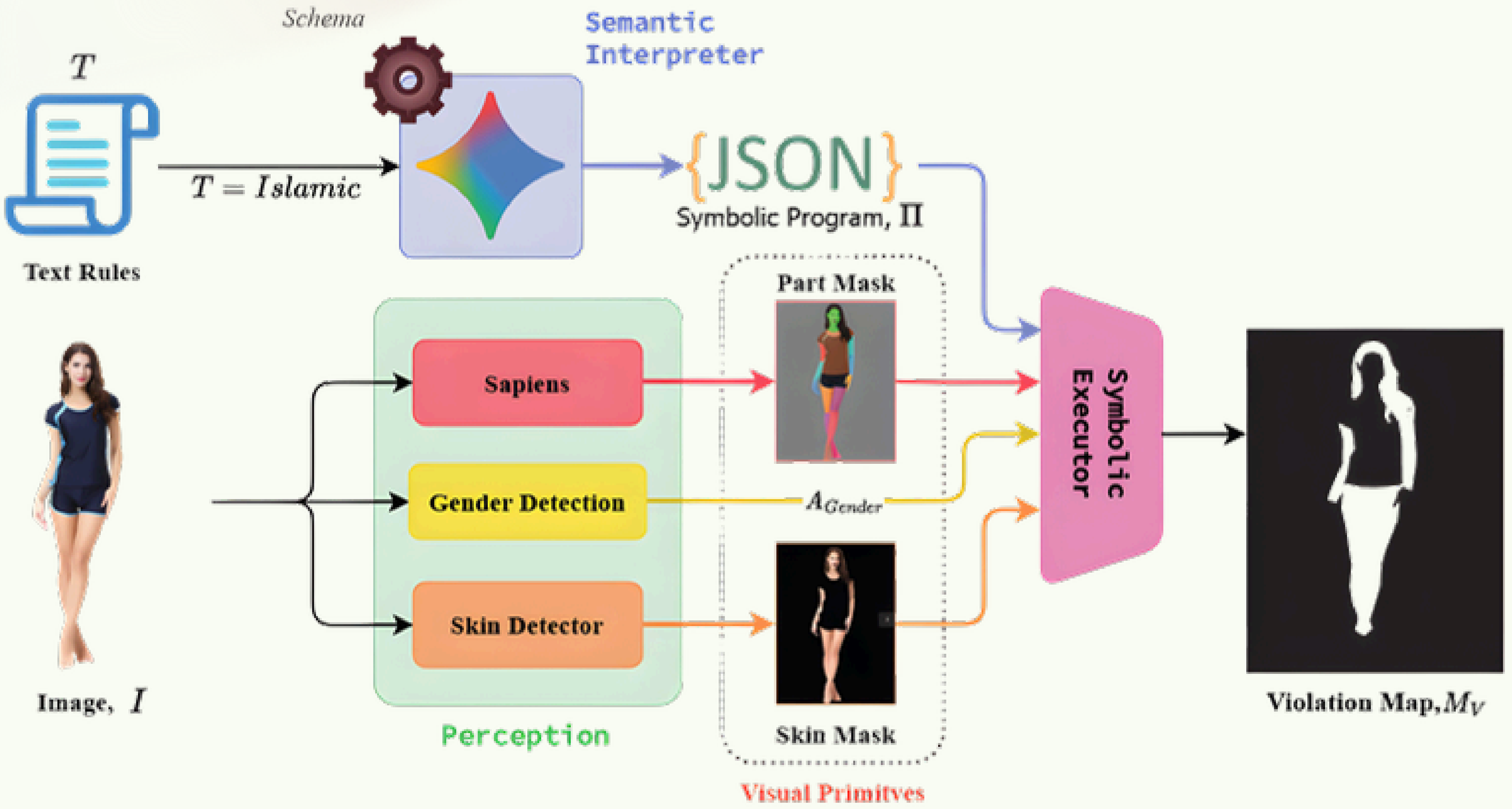
Decouple

LANGUAGE understanding

from

precise VISION tasks

Neuro-Symbolic Framework (LaGPS)



Neuro-Symbolic Framework (LaGPS)

We don’t “predict” the mask

we “compute” it.

Semantic Interpreter + Example Program

- **Input:** A text rule like “Women must cover their hair, neck, and arms.”
- **Processing:** The LLM reads this rule and turns it into a structured program.
- **Output:** The program clearly shows which body parts should be covered or visible, making the system easy to explain and verify.


```
{
 "Hair": false,
 "Face_Neck": false,
 "Arms": false,
 "Reasoning": "Hair, neck, and arms must  be covered,
" Left_Upper_Leg ": false ,
 " Right_Upper_Leg ": false ,
 "Lower_Legs": true,
 "Torso": false,
 "Reasoning": "Upper legs and torso must be covered; lower legs and feet may be shown."
}
```

The HRS Benchmark (Dataset)

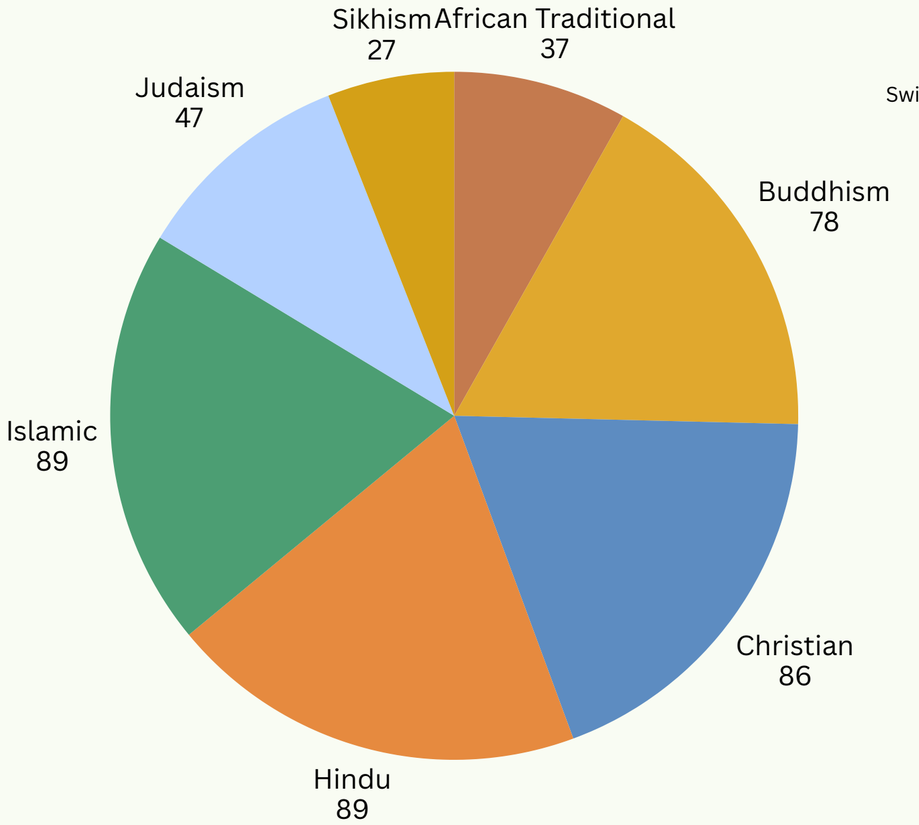
How do we evaluate the accuracy of Policy Segmentation?

- No human dataset exists for precisely grounding policies
- Annotation based on logn-form textual policy

Human-centric Rule-violation Segmentation

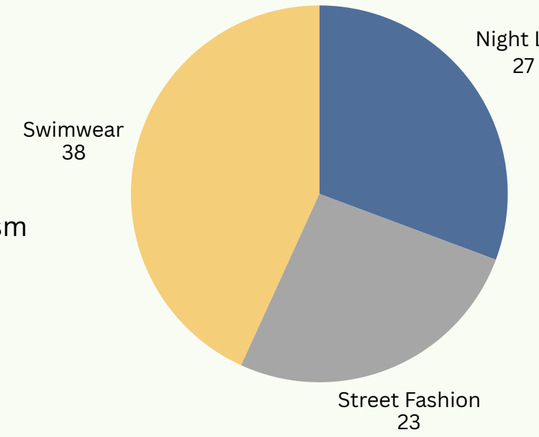
- 1,100 Images
- 16 culturally diverse categories
- Pixel-level annotation of policy violation masks

HRS Benchmark (Distribution)



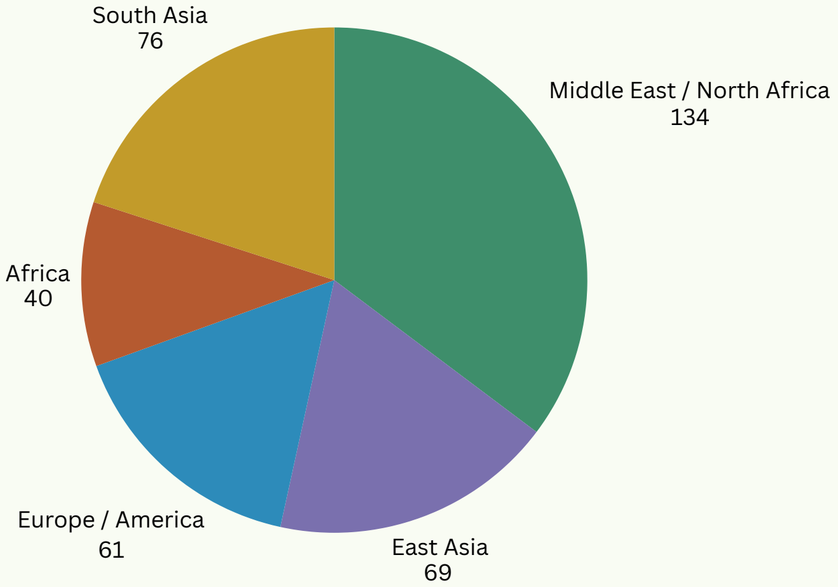
Religion 453

Modern / Urban 88



General 178

Geography 380



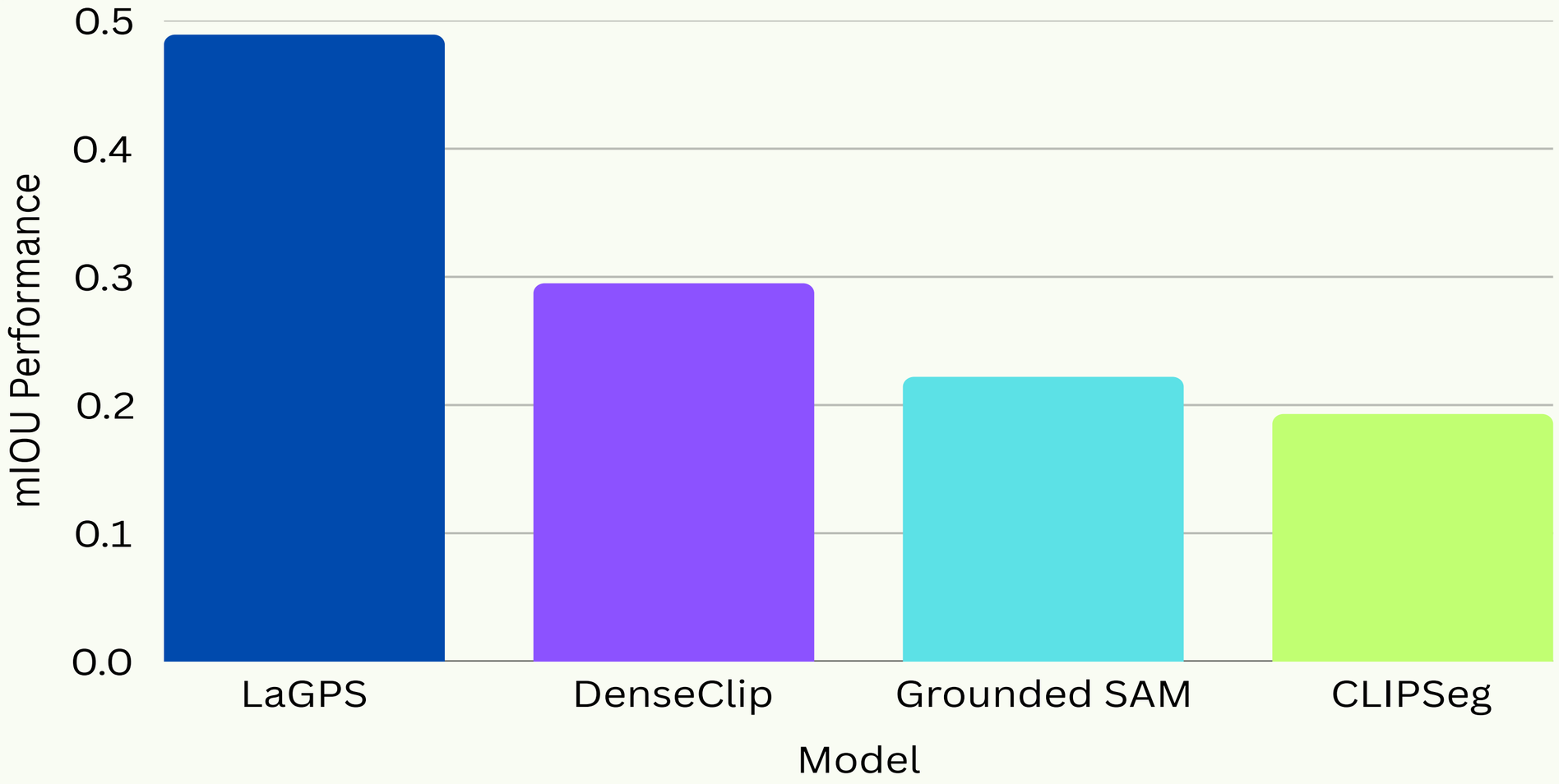
Measuring Segmentation

Rule Adherence Score (RAS)

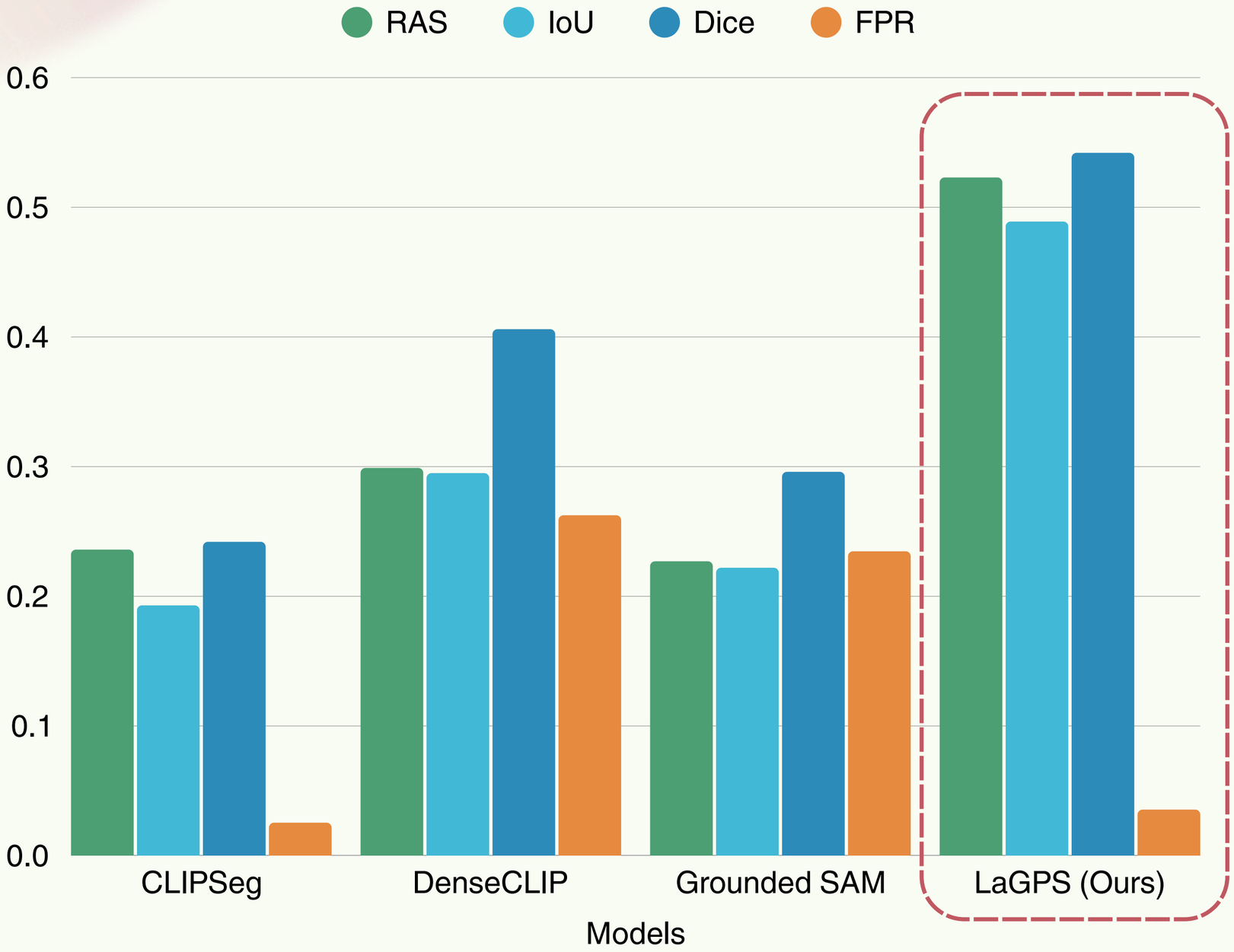
$$RAS = Dice \times (1 - FPR)$$

- Avoid over-segmentation by marking entire people as violations
- Penalize blunt behaviors (False Positives) - “hallucinated violations”

Results (mIoU)



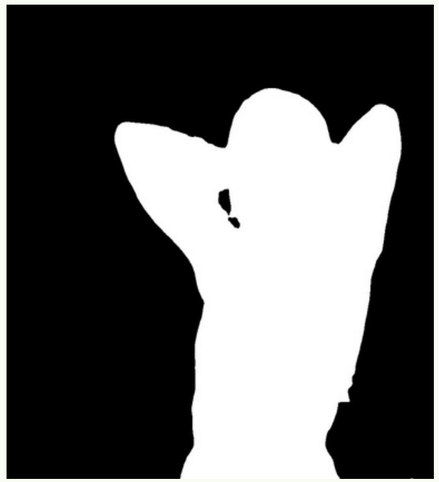
Results


RAS
IoU
Dice
FPR
0.6
0.5
0.4
0.3
0.2
0.1
0.0
CLIPSeg
DenseCLIP
Grounded SAM
LaGPS (Ours)
Models

Qualitative Comparison of Segmentation Results

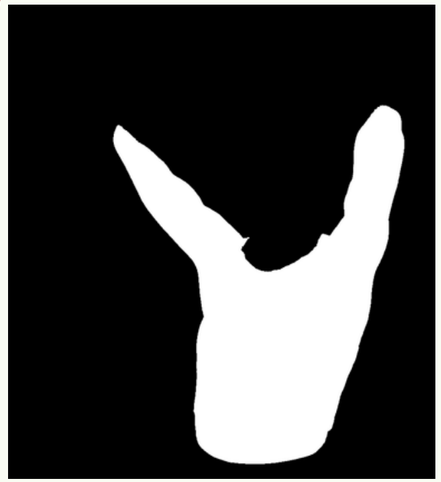
raw image

Clip seg



Dense Clip

Grounded Dino



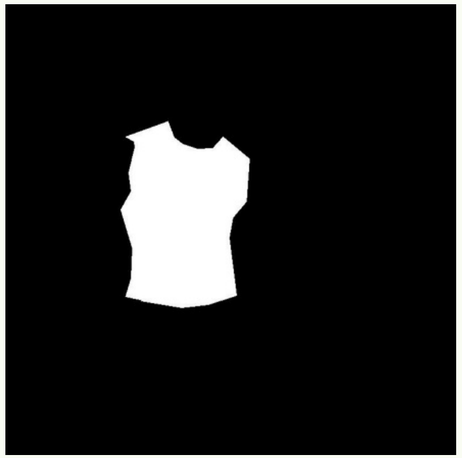
Our Model

raw image

Clip seg

Dense Clip

Grounded Dino



Our Model

Qualitative Comparison of Segmentation Results

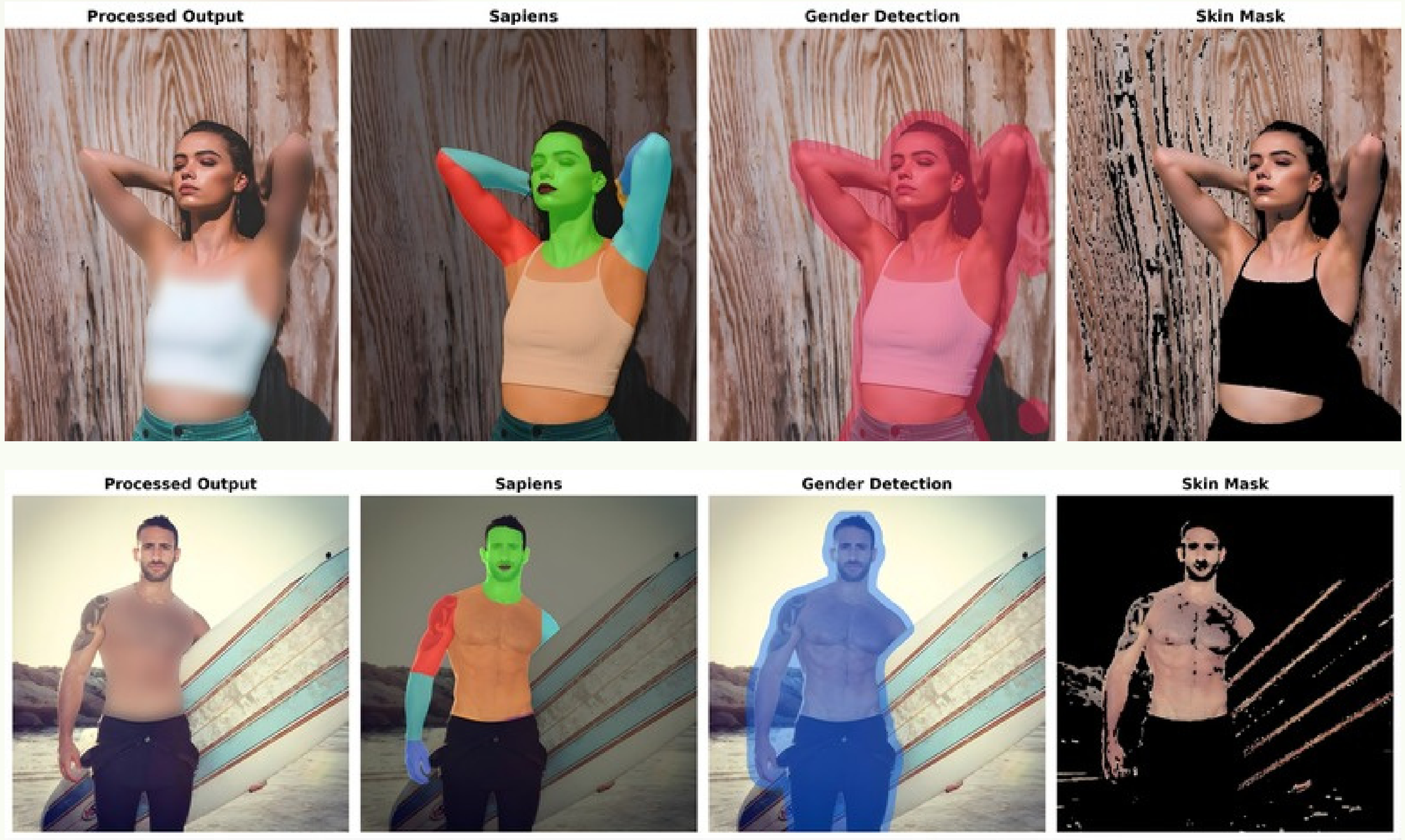
LaGPS isolates only the violating pixels

not the entire person.

Interpretability & Accountability

- Every pixel has a reason trace.
- Failures are classifiable: interpretation, primitives, or execution.

Visualization of Intermediate Steps in LaGPS



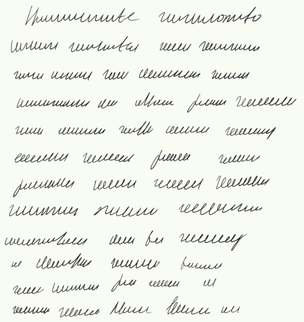
Limitations

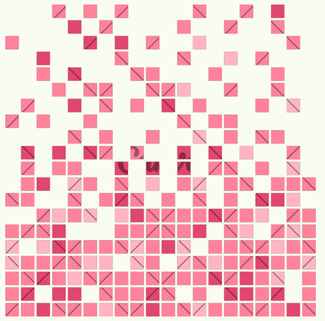
- Dependent on primitive detectors (YOLO, Sapiens, skin)
- Current grammar limited to coverage-style rules
- Ambiguous policies require interpretation

Modular design makes improvements practical and errors diagnosable.

Takeaway

Language → Programs → Pixels





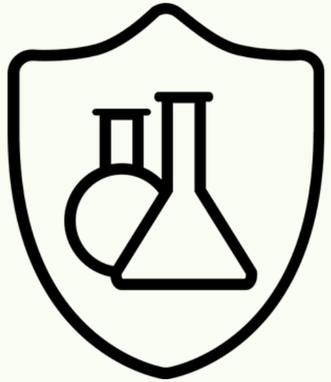
Ethics

Purpose is Transparent and Auditable Visual Grounding of policies

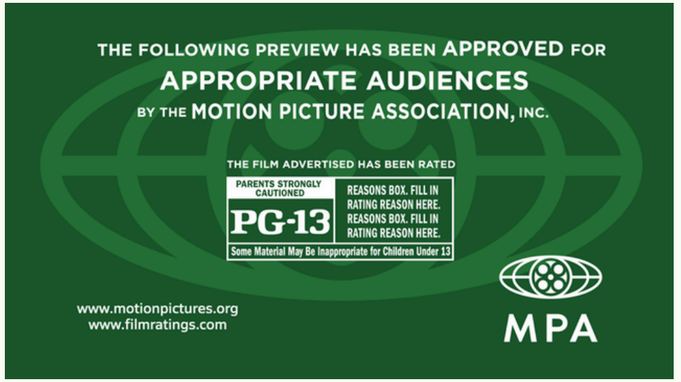
No Auto-enforcement or interpretations of Policy

Beyond Human-centric

PPE Compliance



Lab Safety



Moderation Policies

The architecture applies anywhere

formal textual constraints map to pixels.

THANK YOU

NEURAL INFORMATION PROCESSING SYSTEMS

MUSLIMS IN MACHINE LEARNING

Contacts

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

ALIGARH MUSLIM UNIVERSITY

Northeastern University

1

MOHD HOZAIFA KHAN

MS (by Research)
CSE, IIIT Hyderabad
mohd.hozaifa@research.iiit.ac.in

Actively seeking PhD opportunities.

2

Harsh Awasthi

B.Tech. Computer Engineering
Aligarh Muslim University
harsh.awasthik@gmail.com

3

PRAGATI JAIN

B.Tech. Computer Engineering
Aligarh Muslim University
pragatijain841@gmail.com

4

MOHAMMAD AMMAR

MS Artificial Intelligence,
Northeastern University
ammar.m@northeastern.edu