From Rules to Pixels: A Decoupled Framework for Segmenting Human-Centric Rule Violations

Mohd Hozaifa Khan[1]
mohd.hozaifa@research.iiit.ac.in

Harsh Awasthi[2]
harsh.awasthik@gmail.com

Pragati Jain[2]
pragatijain841@gmail.com

Mohammad Ammar[3]
ammar.m@northeastern.edu

[1]**CSE, IIIT Hyderabad, India**
[2]Computer Engineering, ZHCET, Aligarh Muslim University, India
[3]CSE, Northeastern University, USA

Abstract

We introduce LaGPS, a neuro-symbolic framework that grounds long-form textual rules, such as cultural dress codes, by translating them into deterministic programs for segmentation of rule violations[1]. Existing vision-language models struggle with this task because they cannot parse the compositional logic inherent in human rules. LaGPS overcomes this limitation with a two-stage architecture: a *Semantic Interpreter* that uses a large language model to compile free-form text into a structured program, and a *Symbolic Executor* that runs this program over a set of visual primitives (e.g., per-person body parts, skin masks, etc) to produce precise segmentation masks. To evaluate this setting, we introduce the *Human-Centric Rule-violation Segmentation (HRS)* benchmark for this task, a new $1,100$ image dataset spanning diverse cultural contexts. LaGPS significantly outperforms baselines like CLIPSeg, achieving a $+19.4\%$ absolute mIoU improvement. Our work demonstrates that this decoupled approach creates more transparent, accurate, and auditable systems for language-guided visual reasoning.

1 Introduction

A key challenge for computer vision is moving beyond the recognition of concrete objects towards an understanding of the nuanced, compositional, and cultural contexts inherent in visual data [1]. While modern models excel at identifying discrete objects like a 'cat' or a 'car', they often fail to interpret abstract, human-centric concepts such as what constitutes an 'culturally or contextually conditioned attire requirement' [2]. This gap between low-level recognition and high-level understanding limits the development of AI that can reliably operate in the complexities of human environments [3].

Grounding text-based rules in visual data is a canonical problem that highlights this limitation [4]. Interpreting a seemingly simple rule like "Shoulders and knees must be covered" demands a cascade of reasoning that is trivial for humans but challenging for standard models [5, 6]. This task requires identifying semantic body parts, understanding their state (e.g., 'covered' vs. 'uncovered'), and composing these conditions according to a logical specification [6]. Solving this problem enables transparent and auditable systems for understanding how culturally-situated guidelines are visually

[1]Here, "violation" is used in a strictly technical sense to denote pixels where a *user-specified* visual condition is not met; it carries no moral, cultural, or legal implication.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop.

interpreted by vision-language models. This transparency supports human-in-the-loop review when policy meaning varies across communities and contexts.

Current end-to-end Vision-Language Models (VLMs) struggle with this task due to a fundamental **"semantic-symbolic gap"**: a disconnect between their ability to parse long-form language and the skill needed to execute precise, logical operations on visual evidence [7]. Monolithic models like CLIPSeg [8] attempt a direct mapping from text to pixels, but are not designed to decompose and systematically evaluate the compositional logic required by a rule like "legs must be covered below the knee but above the ankle" [9]. Consequently, they often produce imprecise segmentations that fail to respect logical boundaries [10].

We introduce LaGPS (Language-Guided Policy Segmentation), a neuro-symbolic framework that bridges this gap by compiling natural language into an executable symbolic program. Our contributions are:

- A **Semantic Interpreter**, an LLM-based module that translates unstructured textual rules into a structured, symbolic program defining logical violation[2] conditions.
- A deterministic **Symbolic Executor** that executes this program over primitives (like body parts, gender, etc) to produce precise violation segmentations.
- The **Human-Centric Rule-violation Segmentation (HRS) benchmark**, a new dataset of $1,100$ images with rich categories designed to evaluate nuanced and diverse rule grounding.

2 Related Work

Referential Grounding. The dominant paradigm in vision-language research is referential grounding, which uses implicit feature matching to locate objects described in text [11–14, 8, 15]. These models (CLIPSeg [8], LSeg [15], DenseCLIP [16]) employ contrastive learning or detector-SAM pipelines [14] that conflate semantic interpretation with visual execution [17]. While effective for object identification, this architectural coupling is unsuited for parsing compositional logic and prevents diagnosing error sources, creating a "semantic-symbolic gap" [18].

LLM Planners. A more recent class of models uses LLMs as planners to execute sequences of imperative commands via vision tools [19]. ViperGPT [19] generates Python code that orchestrates API calls for multi-step visual reasoning [20]. These systems are designed for *command execution*, translating user instructions into sequential tool invocations. LaGPS targets a different objective: *policy interpretation for vision tasks*, where the goal is to specify *what constitutes compliance* rather than *how to execute a query*. This distinction enables reasoning over conjunctive/disjunctive constraints (e.g., "neck `AND` hands covered `OR` full-body garment") that are difficult to express as linear action sequences [21, 22].

Neuro-Symbolic Synthesis. To bridge the semantic-symbolic gap, neuro-symbolic synthesis leverages LLMs as compilers that translate natural language into structured programs for deterministic execution [23, 24]. Historically constrained by hand-crafted logical vocabularies, modern implementations generate *declarative programs* for precise evaluation [25], contrasting with the *imperative scripts* of planners or the *opaque feature maps* of grounding models [26]. LaGPS extends this trajectory by explicitly decoupling semantic interpretation (Semantic Interpreter) from visual execution (Symbolic Executor), enabling two capabilities: (1) fine-grained, body-part-level policy specification through a domain-specific language; and (2) diagnostic failure analysis that traces errors to specific components, as demonstrated in our experiments.

LaGPS generates declarative programs for policy interpretation, executed over visual primitives, enabling both compositional logic and diagnostic transparency. This contrasts with grounding models that output *opaque pixel maps* or planners that output imperative scripts for command execution.

3 The LaGPS Framework

Our LaGPS framework grounds textual rules in multi-person scenes through a modular, two-stage pipeline, as illustrated in Figure 1. The pipeline begins by extracting a set of foundational visual

[2]Here and throughout, we use "violation" to refer only to cases where the visual evidence does not satisfy the stated condition, without implying judgment or enforcement.

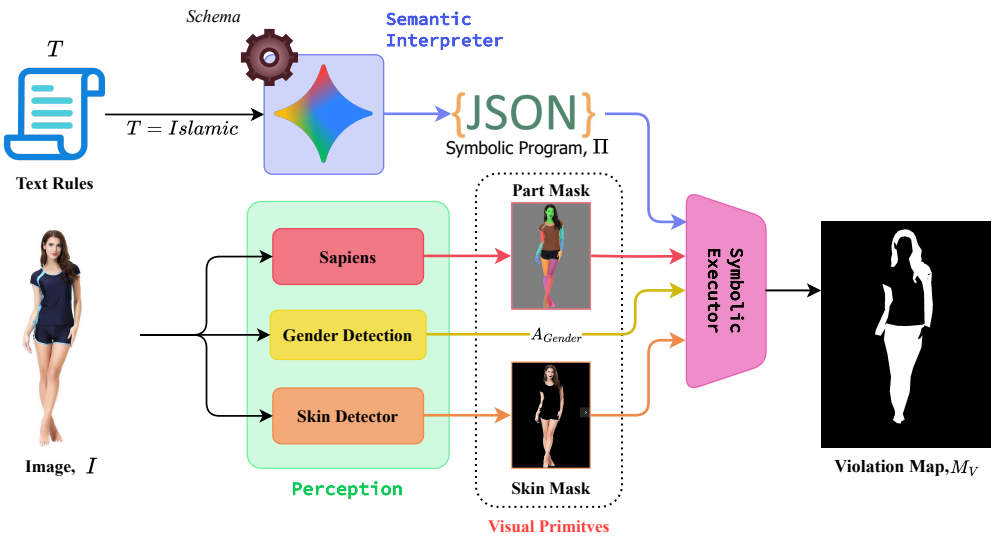

Figure 1: **The LaGPS neuro-symbolic architecture.** (Top) A textual rule is compiled by the Semantic Interpreter into a JSON program. (Bottom) Visual primitives, including a Part Mask ($M_{\texttt{part}}$) and Skin Mask ($M_{\texttt{skin}}$), are extracted from the input image. The Symbolic Executor then evaluates the program Π against the primitives to produce the final Violation Map, M_V.

primitives (Sec. 3.1) from the input image (e.g., body parts). Next, a Semantic Interpreter translates the textual rule into a structured, symbolic program (Sec. 3.2). Finally, a deterministic Symbolic Executor evaluates this program over the visual primitives to produce the violation mask (Sec. 3.3).

3.1 Foundational Visual Primitives

The vocabulary for our Symbolic Executor is constructed from four visual primitives: (1) **per-person instance masks** from a YOLOv8n-seg [27] model for disambiguation; (2) **fine-grained anatomical part masks** (e.g., `face`, `left_knee`) from the Sapiens[3] model[28]; (3) a **global skin mask** generated using game-theoretic approach in YCbCr thresholding [29]; and (4) a **gender classification**[4] for each person. These primitives are extracted for the entire image and then associated with each person instance, enabling person-specific rule evaluation. We provide qualitative examples in Appendix D

3.2 The Semantic Interpreter

The core of our framework is the Semantic Interpreter, which uses a Large Language Model (LLM) to translate the free-form text rule T into a structured, symbolic program Π. By using Structured Decoding, we constrain the LLM to act as a reliable language compiler, outputting a strict JSON schema. The program Π is an "allow list", where a `False` value for a body part indicates that its visibility is a violation. The schema also requires the LLM to output a `reasoning` string for auditability. An example program is provided in Appendix C.3. This process may involve the LLM completing or refining the rule through both literal translation and common-sense inference, while also resolving any gender-specific logic.

3.3 The Symbolic Executor: Rule-based Mask Composition

The Symbolic Executor constructs a violation mask M_V by executing the program Π over the visual primitives. The final mask is the aggregation of all violations across all detected persons, formulated as: $M_V = \bigcup_i \bigcup_{j:\pi_j=\texttt{false}} \left(P_i \cap M_{\texttt{part},j} \cap M_{\texttt{skin}}\right)$ for all person indices i and part indices j where π_j (condition for part j) within Π evaluates to `false` for the attributes of person i. This formulation

[3] We use `sapiens-seg-1B` variant.

[4] "Gender" here refers to classification inferred from visual attribute; see Ethics in Appendix B

ensures every pixel in the final mask can be traced back to a specific rule and a set of visual primitives, making the system's reasoning transparent and auditable.

4 Experiments

4.1 The Human-Centric Rule-violation Segmentation (HRS) benchmark

Motivation. A primary obstacle to developing models for subjective visual reasoning is the lack of appropriate evaluation benchmarks. Existing segmentation datasets are annotated with a fixed vocabulary of concrete objects and cannot be used to evaluate a model's ability to ground compositional, subjective, and culturally-nuanced rules. To address this gap, we created the Human-Centric Rule-violation Segmentation (HRS) benchmark. **Curation.** The ground-truth masks were created by three annotators using a two-stage process with specific guidelines, detailed in Appendix A.2. **Ethics.** For ethical considerations in our work refer to Appendix B

Benchmark Diversity. For evaluation, we introduce the HRS benchmark, which contains *1,100* diverse images organized into a hierarchy of 4 primary and 16 sub-categories. This hierarchy enables testing across a wide range of cultural rules and visual contexts (e.g., Islam, Swimwear, East Asia). The categories include *Religion* (454 images, e.g., Islam, Hinduism), *Geography* (380, e.g., Middle East, East Asia), *Urban* (88, e.g., Swimwear), and *General* (178). A complete breakdown of the dataset hierarchy and image counts is provided in Appendix A.1.

4.2 Experimental Setup

Baselines and Metrics. We evaluate LaGPS against 3 strong baselines: CLIPSeg [8], DenseCLIP [16], and Grounded SAM [14]. For each baseline, we provide the entire textual rule as the input prompt, testing their out-of-the-box compositional reasoning capabilities. Performance is measured using standard segmentation metrics (IoU, Dice [30]) and boundary alignment (Hausdorff Distance) [31]. To specifically penalize the imprecise over-segmentation common in this task, we introduce the *Rule Adherence Score (RAS)* defined as: $\text{RAS} = \text{Dice} \cdot (1 - \text{FPR})$, where the False Positive Rate (FPR) measures the fraction of background pixels incorrectly segmented as a violation. A high RAS score requires both accurate segmentation of the target region (high Dice) and correctly ignoring irrelevant areas (low FPR).

Implementation. We implement LaGPS using publicly available models. We use YOLOv8n-seg [27] for person instance segmentation, the Sapiens [28] model for body part extraction, and Gemma3-4B [32] for gender classification. We use Gemini 2.5 Flash (with zero-shot prompting) as the LLM for our Semantic Interpreter.

4.3 Results and Analysis

Quantitative Results. As shown in Table 1, LaGPS significantly outperforms all baselines across every key metric. Our method achieves a **+19.4%** absolute mIoU improvement over the strongest segmentation baseline, DenseCLIP (0.295). This demonstrates that our neuro-symbolic design, which explicitly compiles language into a symbolic program, is a superior strategy for this task than relying on opaque, end-to-end models.

Table 1: **Quantitative comparison on HRS benchmark**. We report our novel Rule Adherence Score (RAS) and standard metrics. The **best** and second-best results are highlighted. ↑/↓ indicates higher / lower is better. Our model, LaGPS, achieves SOTA performance on almost all metrics.

	Overlap & Precision			Distance & Error	
Model	**RAS ↑**	**IoU ↑**	**Dice ↑**	**Haus. ↓**	**FPR ↓**
CLIPSeg [8]	0.236	0.193	0.242	503.77	**0.0254**
DenseCLIP [16]	0.299	0.295	0.406	464.80	0.2625
Grounded SAM [14]	0.227	0.222	0.296	508.47	0.2347
LaGPS (Ours)	**0.523**	**0.489**	**0.542**	**288.42**	0.0355

LaGPS achieves an RAS of **0.523**, more than $1.5\times$ the score of the next-best method, revealing its high precision. In contrast, the architectural flaws of monolithic models are exposed by their high FPRs; they produce coarse, over-segmented masks that fail to respect logical boundaries. This precision is further confirmed by the Hausdorff distance, where LaGPS is the only model to score below 300. The per-category analysis clarifies this, showing the model excels in high-signal categories like `urban-swimwear` (IoU 0.668) but degrades when primitives fail. Figure 4 in the Appendix D provides a clear qualitative illustration. We can diagnose two failure modes: *aggressive*, where complex attire (`religion-islamic`, IoU 0.349) confuses the primitives and causes a high FPR (0.069); and *passive*, where poor lighting in `urban-nightlife` (IoU 0.382) starves primitives of evidence, leading to high false negatives (FPR: 0.013). This diagnostic ability confirms that our neuro-symbolic design is not only effective but also highly interpretable.

5 Conclusion

In this work, we introduced LaGPS, a neuro-symbolic framework demonstrating that decoupling semantic interpretation from symbolic visual execution enables transparent and precise grounding of compositional language in visual data. By compiling free-form rules into executable programs over visual primitives, LaGPS achieves a $+19.4\%$ absolute mIoU improvement over strong end-to-end baselines on the HRS benchmark. This architecture produces outputs that are *auditable and directly traceable* to the rule logic and visual primitives, enabling diagnostic failure analysis (distinguishing primitive errors from semantic errors) that remains opaque in monolithic models.

Limitations and Future Work. The approach depends on visual primitive extractors, and errors propagate even when the semantic program is correct. Improving primitive robustness through multi-modal learning and expanding the symbolic grammar to richer visual attributes (texture, material, temporal dynamics) are promising directions. Integrating self-correcting mechanisms that exploit programmatic structure to detect and mitigate primitive failures would further strengthen the framework's reliability.

Responsible Use. LaGPS is not intended for compliance monitoring or enforcement. Its purpose is transparent translation of *user-specified* rules towards fine-grained visual reasoning tasks, allowing inspection, contestation, and revision before real-world use. This transparency stems from architectural design: the generated program and execution trace provide a human-readable audit trail. A public release will follow in an extended version with an acceptable-use license prohibiting institutional enforcement and requiring human-in-the-loop review.

Acknowledgements

We thank Umaid Iqbal and Mohd Ammad Rehman for their valuable contributions during the early ideation phase and for insightful discussions that shaped the direction of this work. We also thank the reviewers for their constructive feedback that strengthened this paper.

References

[1] Xuan Wang and Zhigang Zhu. Context understanding in computer vision: A survey. *Comput. Vis. Image Underst.*, 229(C), March 2023. ISSN 1077-3142. doi: 10.1016/j.cviu.2023.103646. URL https://doi.org/10.1016/j.cviu.2023.103646.

[2] Gowreesh Mago, Pascal Mettes, and Stevan Rudinac. Looking beyond the obvious: A survey on abstract concept recognition for video understanding, 2025. URL https://arxiv.org/abs/2508.20765.

[3] Netta Ollikka, Amro Kamal Mohamed Abbas, Andrea Perin, Markku Kilpeläinen, and Stephane Deny. A comparison between humans and AI at recognizing objects in unusual poses. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=yzbAFf8vd5.

[4] D. Rigoni, L. Parolari, L. Serafini, A. Sperduti, and L. Ballan. Weakly-supervised visual-textual grounding with semantic prior refinement. In *Proceedings of the British Machine Vision Conference*, London, UK, 2023. doi: 10.48550/arXiv.2305.10913.

[5] Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, and Wenguan Wang. LOGICZSL: Exploring logic-induced representation for compositional zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[6] Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, Ranjay Krishna, Jiajun Wu, and Hamid Rezatofighi. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*, 2025.

[7] Yun Xing, Jian Kang, Aoran Xiao, Jiahao Nie, Ling Shao, and Shijian Lu. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[8] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, June 2022.

[9] Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! *arXiv preprint arXiv:2503.08723*, 2023.

[10] Soopil Kim, Sion An, Philip Chikontwe, Myeongkyun Kang, Ehsan Adeli, Kilian M. Pohl, and Sang Hyun Park. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20143–20151. AAAI Press, 2024.

[11] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey, 2024. URL https://arxiv.org/abs/2412.20206.

[12] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023.

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[14] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. URL https://arxiv.org/abs/2401.14159.

[15] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ArXiv*, abs/2201.03546, 2022. URL https://api.semanticscholar.org/CorpusID:245836975.

[16] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Sever Topan, David Rolnick, and Xujie Si. Techniques for symbol grounding with satnet. In *Advances in Neural Information Processing Systems*, volume 34, pages 28489–28501, 2021.

[18] Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Q5RYn6jagC.

[19] D'idac Sur'is, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11854–11864, 2023. URL https://api.semanticscholar.org/CorpusID:257505358.

[20] Sukai Huang, Nir Lipovetzky, and Trevor Cohn. Planning in the dark: Llm-symbolic planning pipeline without experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14526–14534. AAAI Press, 2025.

[21] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Language models as code executors: Limitations and challenges. In *Proceedings of the International Conference on Machine Learning*, pages 2834–2848, 2023.

[22] Yifei Li, Jiaming Song, and Stefano Ermon. Language-grounded policy learning for complex tasks. *Advances in Neural Information Processing Systems*, 36:78945–78958, 2023.

[23] Adam Stein, Aaditya Naik, Neelay Velingker, Mayur Naik, and Eric Wong. Neuro-symbolic programming in the age of foundation models: Pitfalls and opportunities. *Proceedings of Machine Learning Research*, vvv:1–18, 2025.

[24] Shuoming Zhang, Jiacheng Zhao, Zheng Wang, Yunji Chen, Xiaobing Feng, and Huimin Cui. Lego-compiler: Enhancing neural compilation through translation composability. *arXiv preprint arXiv:2505.20356*, 2025.

[25] Hossein Rajaby Faghihi, Aliakbar Nafar, Andrzej Uszok, Hamid Karimian, and Parisa Kordjamshidi. Prompt2demodel: Declarative neuro-symbolic modeling with natural language. In *Neural-Symbolic Learning and Reasoning: 18th International Conference, NeSy 2024, Barcelona, Spain, September 9–12, 2024, Proceedings, Part II*, page 315–327, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-71169-5. doi: 10.1007/978-3-031-71170-1_25. URL https://doi.org/10.1007/978-3-031-71170-1_25.

[26] Tilman Hinnerichs, Robin Manhaeve, Giuseppe Marra, and Sebastijan Dumancic. Declarative design of neural predicates in neuro-symbolic systems, 2025. URL https://arxiv.org/abs/2405.09521.

[27] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. doi: 10.1109/ADICS58448.2024.10533619.

[28] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IV*, page 206–228, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73234-8. doi: 10.1007/978-3-031-73235-5_12. URL https://doi.org/10.1007/978-3-031-73235-5_12.

[29] Djamila Dahmani, Mehdi Cheref, and Slimane Larabi. Zero-sum game theory model for segmenting skin regions. *Image and Vision Computing*, 99:103925, 2020. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2020.103925. URL https://www.sciencedirect.com/science/article/pii/S0262885620300573.

[30] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing, 2017. ISBN 9783319675589. doi: 10.1007/978-3-319-67558-9_28. URL http://dx.doi.org/10.1007/978-3-319-67558-9_28.

[31] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. doi: 10.1109/34.232073.

[32] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr,

Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

[33] Nikita Manovich, Boris Sekachev, Maxim Zhiltsov, Dmitriy Krapivin, Maksim Kovalev, Eugene Khvedchenya, Vadim Vsesviatskyi, Artem Glushchenko, Mikhail Krivoshein, Fedor Zav'ialov, Pavel Rotmistrov, Egor Potemkin, and Kirill Makhov. opencv/cvat: v1.1.0, August 2020. URL `https://doi.org/10.5281/zenodo.4009388`.

A HRS Benchmark

A.1 HRS Benchmark Composition

The Human-Centric Rule-violation Segmentation (HRS) benchmark is designed to provide comprehensive and diverse coverage for evaluating nuanced rule grounding. Figure 2 visualizes the hierarchical composition of the benchmark, breaking down the 1,100 total images into 4 primary categories and 16 sub-categories. This ensures that models are tested against a wide spectrum of visual contexts, cultural norms, and demographic representations, making it a robust benchmark for future research in this domain.

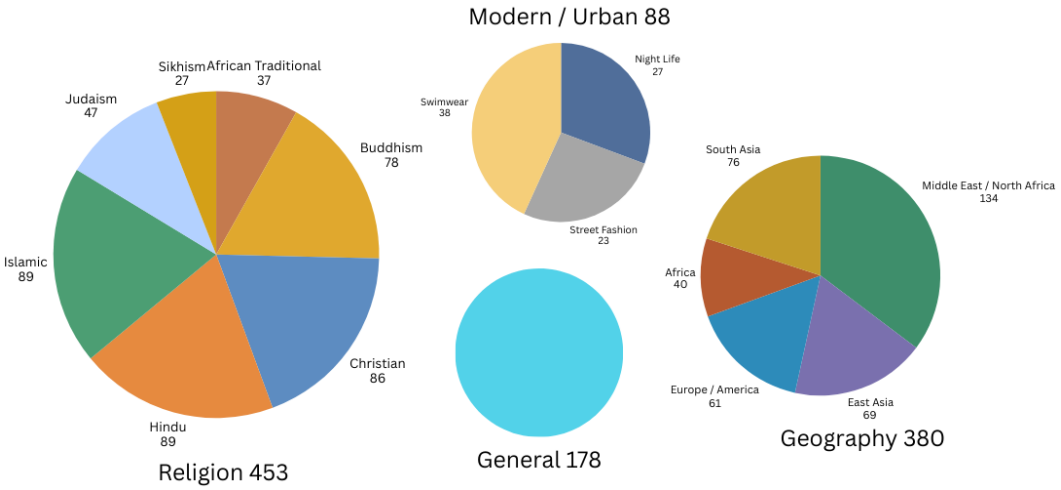

Figure 2: **Hierarchical distribution of the HRS benchmark.** The 1,100 images in the benchmark are organized into four primary categories. The `Religion` and `Geography` categories are further broken down into diverse sub-categories to ensure broad cultural and regional representation. The `Modern/Urban` category focuses on contemporary Western contexts, while the `General` category includes a wide array of miscellaneous scenes. This multi-level, diverse composition is designed to rigorously test a model's ability to ground subjective rules across varied domains.

A.2 Annotation Process

As shown in Figure 3, the ground-truth masks for the HRS benchmark were generated by a team of three annotators according to a detailed protocol. For each sub-category, annotators were provided with specific guidelines and illustrative examples of rule violations.

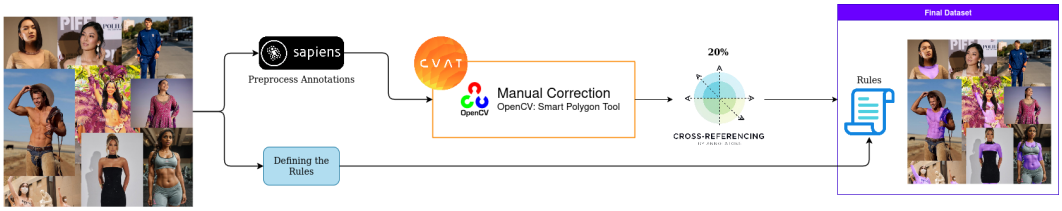

Figure 3: **The HRS data annotation pipeline**. The process proceeds in four stages: (1) Annotators are provided with a rule and a set of candidate images. (2) Initial body part proposals are automatically generated using the Sapiens model to accelerate the workflow. (3) Using the CVAT[33] annotation tool, annotators perform manual correction to create pixel-perfect ground-truth violation masks. (4) Finally, a 20% random subset of each group is reviewed by a second annotator to ensure high inter-annotator agreement and quality.

To ensure consistency, all annotators were provided with a detailed protocol document. For each image and its associated textual rule, annotators were instructed to:

1. Identify all persons to whom the rule applies.
2. For each person, identify the specific body part(s) that violate the rule. For instance, for the rule “Shoulders and knees must be covered," only the exposed skin on the shoulder and knee regions should be segmented.
3. Produce a pixel-perfect segmentation mask for only the violating regions. Areas covered by clothing, even if they are part of a violating body part (e.g., “the part of the knee covered by shorts"), should not be included in the mask.
4. If the rule is ambiguous, adhere to a conservative interpretation provided in a supplementary guide.

To ensure **inter-annotator consistency** (IAA), we employed a two-stage quality assurance process:

1. An initial set of annotations was created by a single annotator.
2. A 20% subset of these annotations was then reviewed by a second annotator to ensure consistency.

B Ethical Statement

B.1 Scope and Intended Use

LaGPS is a research prototype that *grounds textual policies into interpretable visual reasoning steps*. It is **not** intended for enforcement or surveillance. The system applies attribute-conditioned rules **only when the nuance is relevant in the text**. The goal of LaGPS is to build **auditable, transparent systems** that faithfully expose how a *user-supplied* policy is visually interpreted and grounded, rather than to judge or prescribe compliance.

B.2 HRS Dataset

Images in HRS were collected from public web sources under permissive licenses. To enable broad evaluation of appearance-related visual reasoning, the benchmark intentionally includes a range of attire, including *explicit content* for stress-testing segmentation across diverse cultural and contextual settings. Because the images are public-source, *societal biases may persist* and performance can vary across demographic groups despite our effort to create a diverse hierarchy of categories.

B.3 Bias and Misuse

The policies in HRS are sourced from publicly documented cultural and contextual dress guidelines; we **do not assert whether these policies should or should not be followed**. LaGPS reproduces the logical structure of the rules provided to it – it does not correct or endorse the policy itself. We recognize the potential for misuse, particularly automated enforcement of discriminatory dress codes. *Users and the contexts they define* bear **full ethical responsibility** for the policies they supply.

B.4 Non-Enforcement and Community Protection

LaGPS is **not** intended for compliance monitoring or disciplinary enforcement. The system does not judge whether a cultural or religious guideline should be followed; it only makes the *interpretation* of a user-provided rule explicit and auditable. Because the HRS dataset includes culturally-situated attire and may carry social sensitivity, we will release the dataset and code **after** further community review and refinement in a forthcoming extended version. The release will be governed by an **acceptable-use license** prohibiting deployment in institutional or governmental enforcement settings, workplace or school monitoring workflows, and requiring human-in-the-loop review. This staged release is intended to ensure **responsible access, contextual guidance, and safeguards** before broader dissemination.

B.5 Gender References

When a rule mentions gender, LaGPS treats the term as a presentation-based linguistic cue extracted from the wording of the rule and usually applies it only when the cues are relevant in the policy.

B.6 Transparency / Auditability

Every output of LaGPS includes: (1) a structured, machine-readable program describing the logical interpretation of the rule (reasoning chain), and (2) the visual primitives used in its execution. This enables researchers or affected individuals to **inspect, contest, revise or reject** the interpretation **prior to any real-world deployment**.

C Implementation Details

C.1 Structured Decoding for Program Generation

The Semantic Interpreter ensures the generated symbolic program Π adheres to a strict schema using structured decoding. A Pydantic schema defines the program's structure, specifying required fields for body parts and their boolean values (e.g., covered or uncovered). This constrains the LLM's output to be syntactically and semantically valid, enhancing reliability and simplifying execution by the Symbolic Executor.

C.2 Reproducibility Details.

Interpreter. We use `Gemini 2.5 Flash` in zero-shot mode (temperature=0.0, top-p=0.9). The exact prompt templates and the Pydantic schema used for structured decoding are provided in Appendix C.3. Decoding is performed via a schema validator; failed generations trigger up to two re-prompts with a consistency hint.

Visual Primitives. Person masks use YOLOv8n-seg; part masks use Sapiens-1B [28]; skin masks are computed via YCbCr thresholds [29]. When the rule text explicitly differentiates by gender, we infer *gender presentation* by prompting for each person (Gemma3-4B) and classifying presentation via a constrained LLM decision set; otherwise, this attribute is ignored.

Execution and Evaluation. The symbolic executor operates over binary masks using set operations. We evaluate Dice, IoU, FPR, and RAS. All experiments ran on a single Nvidia-T4 and Cuda 12.8.

Release. We will release code and data under an acceptable-use license that prohibits surveillance or coercive enforcement, with a documented takedown mechanism.

C.3 Example of a Generated Symbolic Program

This section provides an example of the symbolic program Π generated by our Semantic Interpreter for the textual rule: "In this culture, women must cover their hair, neck, and arms. Legs below the knee may be shown." Using structured decoding, the LLM is constrained to output a program that conforms to a fixed Pydantic schema, evaluating the rule's impact on all possible body parts. The program uses an "allow list" logic, where a boolean value of `False` indicates the corresponding body part must be covered.


```
{
  "Hair": false,
  "Hair_Reasoning": "Hair must be covered as per cultural rules.",
  "Face_Neck": false,
  "Face_Neck_Reasoning": "The neck must be covered; face visibility is
    not restricted, but the combined region means the neck cannot be
   shown.",
  "Right_Upper_Arm": false,
  "Left_Upper_Arm": false,
  "Right_Lower_Arm": false,
  "Left_Lower_Arm": false,
  "Right_Wrist": false,
  "Left_Wrist": false,
  "Arms_Wrist_Reasoning": "All parts of the arms, including wrists,
   must be covered as per cultural rules.",
  "Left_Upper_Leg": false,
```

```
  "Right_Upper_Leg": false,
  "Left_Lower_Leg": true,
  "Right_Lower_Leg": true,
  "Left_Foot": true,
  "Right_Foot": true,
  "Legs_Feet_Reasoning": "Upper legs must be covered. Lower legs and
   feet may be shown as per cultural rules.",
  "Torso": false,
  "Torso_Reasoning": "Although not explicitly stated, the requirement
   to cover the neck and arms implies that the torso must also be
   fully covered to maintain modesty."
}
```

Listing 1: Example symbolic program Π generated by the Semantic Interpreter using structured decoding.

D Additional Results

D.1 Qualitative Examples of the LaGPS Pipeline

To provide a clear, visual understanding of our framework, Figure 5 illustrates the sequential outputs of the core modules in the LaGPS pipeline for three distinct examples. The figure showcases the intermediate outputs for the foundational visual primitives that our Symbolic Executor uses to execute the final rule-based segmentation.

As demonstrated in Figure 5, the LaGPS pipeline robustly grounds rules across diverse cultural contexts. We highlight several key examples:

- **Row 1 (Hinduism):** LaGPS successfully applies a rule regarding modest swimwear, accurately segmenting the bikini top and bottom.
- **Row 2 (Islam):** The framework correctly grounds a rule requiring the covering of the hair (hijab), producing a precise segmentation of the head and neck region.
- **Row 3 (Orthodox Judaism):** For a rule related to modest dress (tzniut), the model correctly identifies the hair and upper body as the regions of interest.
- **Row 4 (Islam):** This example highlights the system's abstract reasoning. For a male subject under a rule requiring the torso and arms to be covered, LaGPS correctly identifies the entire anatomical region as a violation, even though the subject is shirtless. This shows the model grounds the rule's logic on the body itself, not just on existing clothing.

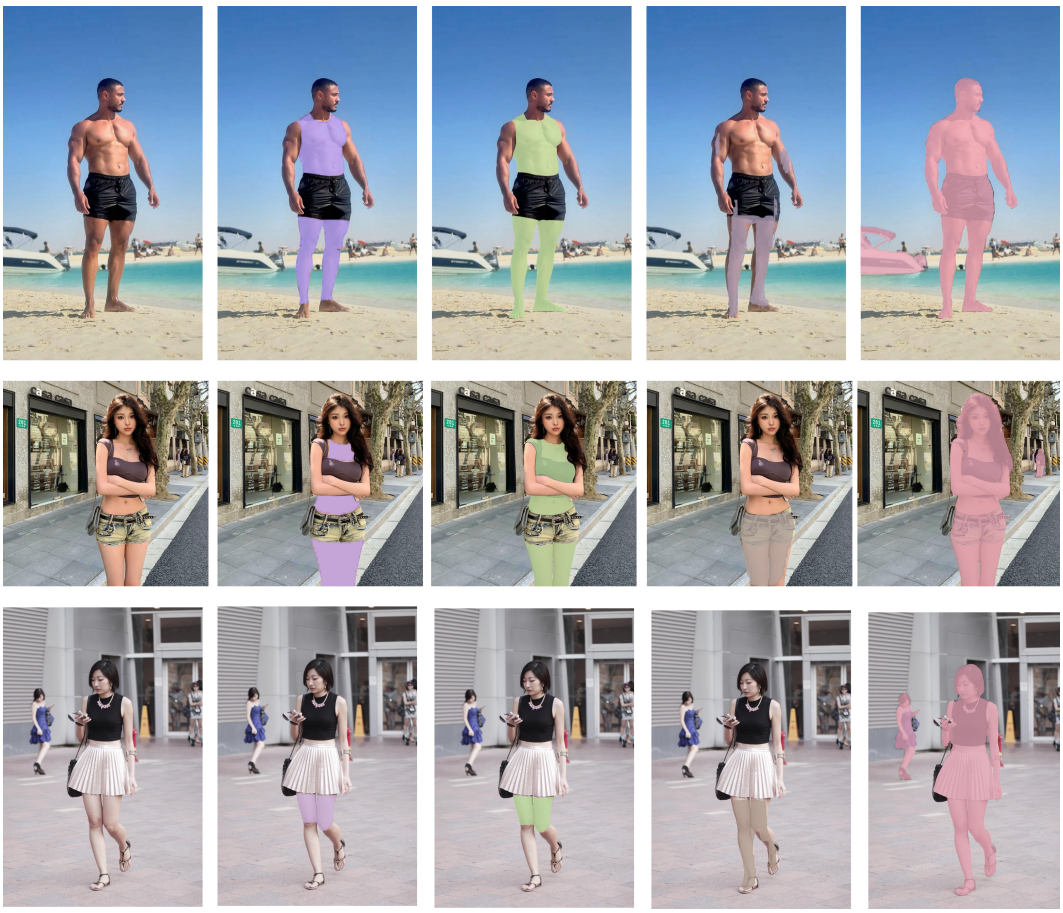

Figure 4: **Qualitative comparison with baselines on the HRS benchmark.** Our model (LaGPS) produces precise and accurate violation masks that closely align with the ground truth. In contrast, the end-to-end models exhibit characteristic failure modes. CLIPSeg often produces low-confidence, imprecise masks that fail to capture sharp boundaries. Grounded SAM consistently over-segments, failing to distinguish between the person as a whole and the specific body parts targeted by the rule (Rows 1-3), and even incorrectly includes background individuals (Row 3). This visually demonstrates the "semantic-symbolic gap", the inability of monolithic models to execute the precise, compositional logic required by the rule.

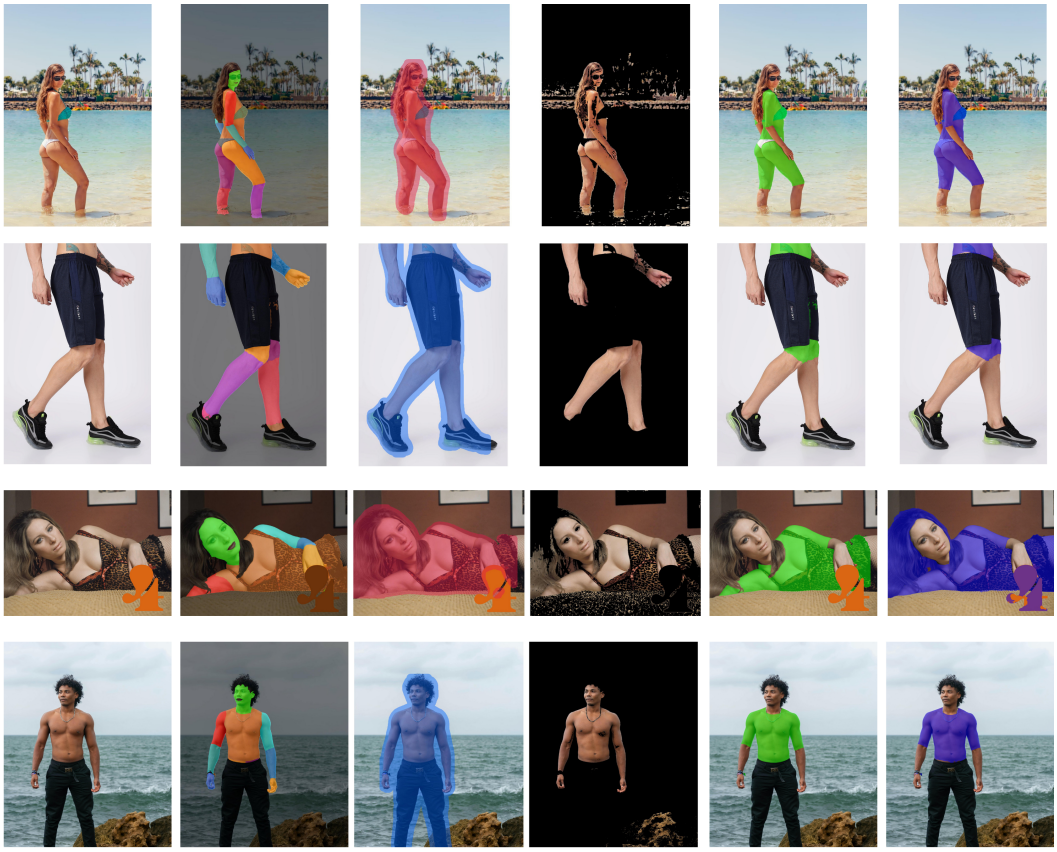

Figure 5: **End-to-end visualization of the LaGPS pipeline on culturally-specific rules.** This figure demonstrates the framework's robustness across diverse religious dress codes by showing key stages: the original image (1), extracted primitives (2, 3), person mask (4), final LaGPS output (5), and ground truth (6).