The *Sesame Street* Archive: a labeled image repository of educational children's television, 1969-2018

Karol Sadkowski[1], Siyuan Guo[2], Chen Yu[3], Sophia Vinci-Booher[1]

[1]Psychology and Human Development, [2]Computer Science, Vanderbilt University; [3]Psychology, University of Texas at Austin

{karol.sadkowski, siyuan.guo, sophia.vinci-booher}@vanderbilt.edu; chen.yu@austin.utexas.edu

SSA web app:

This poster:

CVPR Nashville JUNE 11-15, 2025

1. INTRODUCTION

The *Sesame Street* Archive (SSA) is the first labeled image repository sourced from a children's educational television series.

- Projected to contain 35,000+ film frames from 4,397 episodes of *Sesame Street* aired between 1969–2018.
- Features 4 labeled object categories—`face`, `place`, `word`, and `number`—with striking intra-category variation (e.g., faces of humans, animals, puppets).
- Captures real-world, animated, and imaginative contexts through which children learn foundational literacy and numeracy.
- Spans 6 decades of sociocultural change and evolving broadcast formats (e.g., resolution, color, aspect ratio).
- Supports integrated tasks in face detection, OCR, and scene understanding.

2. RELATED WORKS

Existing datasets rarely foreground educational content, and none do so through children's programming.

- Datasets like ChildPlay [3, 9] and Toybox [10] offer limited annotation depth and only incidental educational content.
- The SSA fills gaps left by task-specific datasets like YLFW [6] (faces), SVHN [7] (numbers), and SUN [12] (places).

3. THE SESAME STREET ARCHIVE DATASET

Object Categories:

- Annotation schemas give object categories qualitative structure (Tab. 1).
- Schemas account for stylization, occlusion, non-canonical forms, and hybrid entities.

Data Collection:

- 343 *Sesame Street* episodes received from GBH [4].
- Frames extracted from episodes at 1fps yielded ~1.09M total frames.
- Frames featuring target categories curated to form a training dataset of 4,832 frames (Tab. 2).

Image Annotation:

- Trained coders annotated the training dataset using CVAT [1].
- New coders consensus-annotated individual object categories (Fig. 1).
- Edge cases resolved with iterative criteria and guiding principles.

Ethics and Copyright:

- Permission granted by Sesame Workshop and the Joan Ganz Cooney Center.
- Dataset to be shared for research under FAIR principles [10]; no commercial release.

Attribute	Values
`representation:`	`real-world; stylized; other`
`species:`	`human; puppet; animal; other`
`race-ethnicity:`	`white; black; asian; native-american; pacific-islander; other`
`age:`	`infant; child; teen; adult; elderly; other`
`orientation:`	`frontal; profile; other`
`camera-angle:`	`forward; downward; upward; other`
`visibility:`	`full; occluded; truncated; occluded-and-truncated; other`
`clarity:`	`clear; blurry; other`

Table 1. Annotation schema for identifying and labeling `face` instances. Annotation schema attributes and values reflect object label configurations made in CVAT.

	Instances		Frames	
Subset	Current	Projected	Current	Projected
`face`	7,214	52,251	3,433	24,865
`place`	1,486	10,763	1,000	7,243
`word`	4,672	33,839	1,359	9,843
`number`	2,333	16,898	1,182	8,561
Total	15,705	113,751	4,832*	35,000*

Table 2. Current and projected compositions of ssa_ subsets based on instance and frame counts. Asterisks indicate total frame counts are lower than the sums of their individual subsets, as object instances of different categories co-occur within frames.

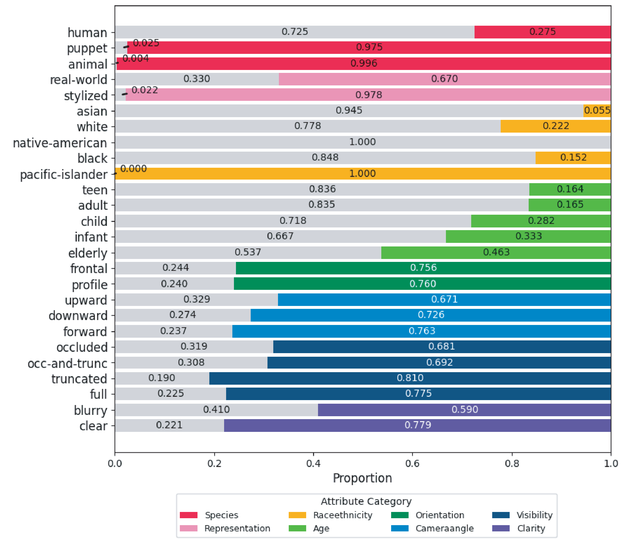
4. ANALYSIS AND RESULTS

Analysis of the `ssa_face` subset highlights remarkable facial diversity in child-centered scenes.

- Of all human `ssa_face` instances, 49% depict children and 31% depict nonhuman entities.
- Educational scenes also offer diverse task semantics, unlike the cluttered scenes commonly used for adult benchmarks.

Current models performing on `ssa_face` underscore a domain bias for adult-centered scenes.

- RetinaFace [2] misses nearly all nonhuman and stylized `ssa_face` instances (e.g., puppets, children's drawings) (Fig. 2).
- RetinaFace & S^{3}FD [12] both exhibit marked precision-recall disparities, suggesting sparse yet precise predictions within child-centered scenes (Tab. 3).



	IoU@0.5			IoU@0.5:0.95
Model	Precision	Recall	AP	mAP
RetinaFace	0.80	0.24	0.94	0.57
S^{3}FD	0.79	0.22	0.95	0.51

Table 3 (left). Performance of RetinaFace and S^{3}FD on the `ssa_face` subset, evaluated at IoU thresholds 0.5 and 0.5:0.95.

Figure 2 (above). Normalized proportions of `ssa_face` instances missed by RetinaFace, evaluated at IoU threshold 0.5. Color-coded segments show proportions missed, while gray segments show proportions detected. Left-side attribute value names that share a common attribute are grouped by color.

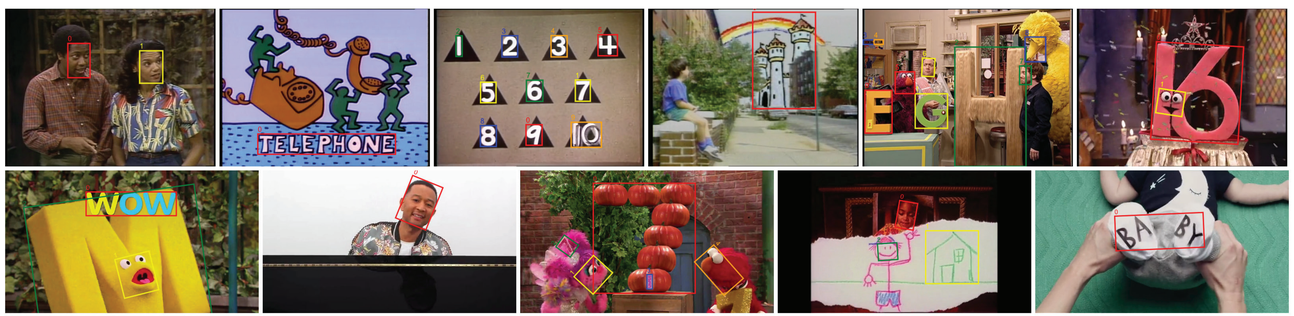


Figure 1. Example Sesame Street film frames in the SSA dataset, each with bounding boxes around faces, words, numbers, and buildings (places). Bottom-row frames have slightly wider aspect ratios, reflecting a historical shift in film production and broadcasting technology.

5. ONGOING AND FUTURE WORK

- Refine annotation schemas, switch to instance segmentation, and enable image captioning.
- Scale to 35,000+ labeled images, prioritizing visually abstract instances and scenes invoking sense of wonder.
- Integrate multimodal data (e.g., eye, cursor tracking) and international *Sesame Street* adaptations.
- Explore the SSA's uses in supporting model domain generalization, grounded in childhood cognition.
- Launch an independent SSA governance consortium and open-source research platform.

REFERENCES

1. CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT). https://www.cvat.ai/, 2023. Accessed: 2025-04-03.
2. Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage Dense Face Localisation in the Wild, 2019. arXiv:1905.00641.
3. Arya Farkhondeh, Samy Tafasca, and Jean-Marc Odobez. ChildPlay-Hand: A Dataset of Hand Manipulations in the Wild, 2024. arXiv:2409.09319.
4. GBH Media Library and Archives. Archives. https://www.wgbh.org/foundation/archives. Accessed: 2025-04-06.
5. Iurii Medvedev, Farhad Shadmand, and Nuno Gonçalves. Young Labeled Faces in the Wild (YLFW): A Dataset for Children Faces Recognition, 2023. arXiv:2301.05776.
6. Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.
7. Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision, 77(1):157–173, 2008.
8. Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. ChildPlay: A New Benchmark for Understanding Children's Gaze Behaviour, 2023. arXiv:2307.01630.
9. Xiaohan Wang, Tengyu Ma, James Ainooson, Seunghwan Cha, Xiaotian Wang, Azhar Molla, and Maithilee Kunda. The Toybox Dataset of Egocentric Visual Object Transformations, 2018. arXiv:1806.06034.
10. Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, and et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1):160018, 2016. Publisher: Nature Publishing Group.
11. Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3485–3492, 2010.
12. Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S^{3}FD: Single Shot Scale-invariant Face Detector, 2017. arXiv:1708.05237.