BIV-Priv-Seg: Locating Private Content in Images Taken by People With Visual Impairments

- Supplementary Materials

Yu-Yun Tseng[1], Tanusree Sharma[2], Lotus Zhang[3], Abigale Stangl[4], Leah Findlater[3], Yang Wang[2] and Danna Gurari[1]

[1] University of Colorado Boulder, [2] University of Illinois at Urbana-Champaign,

[3] University of Washington, [4] Georgia Institute of Technology

Table of contents:

1. Private Object Categories

The 16 categories in our dataset include 14 categories adopted from previous work [4] and 2 additional categories, *pregnancy test box* and *condom packet*. These categories are to complement *pregnancy test* and *condom box* in the previous work [4] since sometimes the pregnancy tests and condoms were photographed outside the carton boxes while other times they were not. Consequently, we differentiate the contents inside a box from the box itself. We also replaced the category *tattoo* with *tattoo sleeve*. This is because of the technical difficulties of annotating tattoo patterns on tattoo sleeves that individuals wore to mimic tattoos; the definition of the area/range of a tattoo pattern is often vague. We instead leverage *tattoo sleeve*, which we define as the sleeve object itself that includes the clothing area where tattoo patterns may be absent.[1]

We note that the category "receipt" already exists in the 100 categories in VizWiz-FewShot. Even though it is indicated in [4] that bills and receipts are considered private objects for BLV, we preserve both categories in VizWiz-FewShot and BIV-Priv-Seg as both are valuable data in the field. We conducted data post-processing to remove the category in BIV-Priv-Seg before model benchmarking to support the validity of our experiments.

[1] In the annotation interface, these special categories are provided with more detailed instructions to prevent confusion in the definition of the categories.

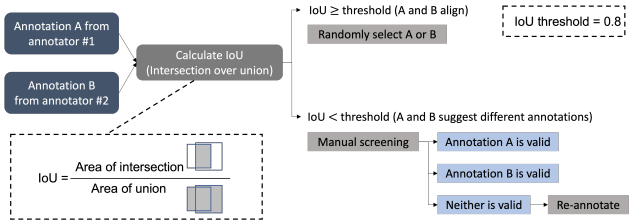


Figure 1. Overview of quality control workflow for generating ground truth instance segmentations using the two instance segmentations collected per object.

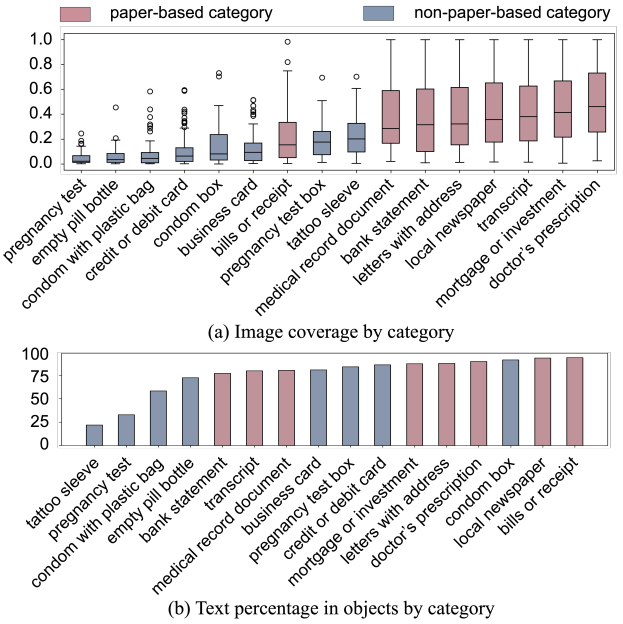
2. Annotation Collection Workflow

For the annotation collection of BIV-Priv-Seg, we adopted the methods proposed in previous work [5]. For a comprehensive quality control, we use the two-annotation quality control approach. An overview of the workflow is illustrated in **Figure 1**.

For the private categories in our dataset, we excluded the additional step of classifying segmented object categories proposed in [5], since the number of categories is relatively smaller, at only 16 category types. In VizWiz-FewShot [5], the 100 categories were separated into 20 categories per batch through the classification task to ensure the quality of the segmentation task.

3. Data Analysis on Paper-Based and Non-Paper-Based Categories

Figure 2(a) illustrates the image coverage distribution of each category in our dataset. We hypothesize that it is more challenging for the BLV group to adjust the camera to a proper distance from an A4 or letter-sized paper resulting in many paper-based categories tending to have higher image coverage than non-paper-based objects.



(a) Image coverage by category

(b) Text percentage in objects by category

Figure 2. Results of category-based dataset analyses on BIV-Priv-Seg. The plots show (a) the proportion of instances with text sorted by frequency of text for some of the categories and (b) the distribution of image coverage sorted by the medians of the distributions.

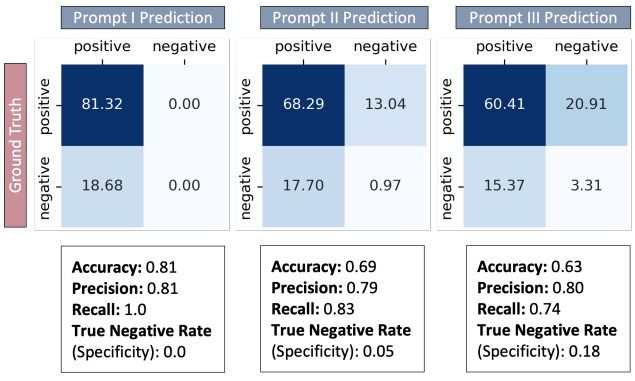


Figure 3. GLaMM privacy classification results of prompts I, II, and III on BIV-Priv-Seg in binary classification metrics. The confusion matrices (TP, FP, TN, and FN) are presented in percentages.

4. Evaluation Results of Mask Level Privacy Classification

The detailed evaluation of the GLaMM privacy classification results is presented and analyzed in the section. First, we show the classification results at the *mask level* for the three prompts in **Figure 3**. While prompt I achieves the highest accuracy, it does not accurately reflect the model's classification capability since all samples in the dataset are classified as private. Across all tested prompts, it is evident that the model struggles to correctly identify negative samples, meaning no target object is present.

Breaking down the 16 categories prompted separately in the prompt I scenario (category-specific scenario), we present the mask-level results of the multi-class classification, including metrics such as confusion matrices in **Figure 4**. The observations align with those from the binary privacy classification: there are generally high recall scores and low true negative (TN) rates. This suggests that while the model has a strong ability to identify positive samples, it struggles significantly with identifying negative samples.

5. Qualitative Results of VLM Benchmarked on BIV-Priv-Seg

Figure 5 shows qualitative examples of GPT-4o coupled with Dall-E2 with the prompt *"Generate a binary mask that segments the {category name}."*. The model failed to generate binary masks for segmentation. **Figure 6** shows qualitative examples CogVLM with prompts I. The model performs poorly by achieving a mAP score at 0.

Figure 7 visualizes the examples of prediction results of GLaMM on our dataset. We compare the predictions of prompt I, prompt II, and prompt III to ground truth. Shown are examples of both images with accurate and inaccurate segmentation results from prompt I. In the examples of accurate predictions from prompt I, prompt II and prompt III, results often struggle to well-segment the private object. However, in examples such as the bank statement and the pregnancy test in **Figure 7**, the explanation of private object leads prompt III to perform better than prompt II.

Following the implementation source[2], we benchmarked the model of GroundingDINO [2] coupled with SAM [1] on our dataset. **Figure 8** visualizes examples of prediction results. Two prompts are used: (1) *"{category name}"* and (2) *"private object"*. Results show a similar trend as those of GlaMM, where the model performs better when given the exact category name, while struggling with the more vague notion of "private object".

[2]https://github.com/luca-medeiros/lang-segment-anything

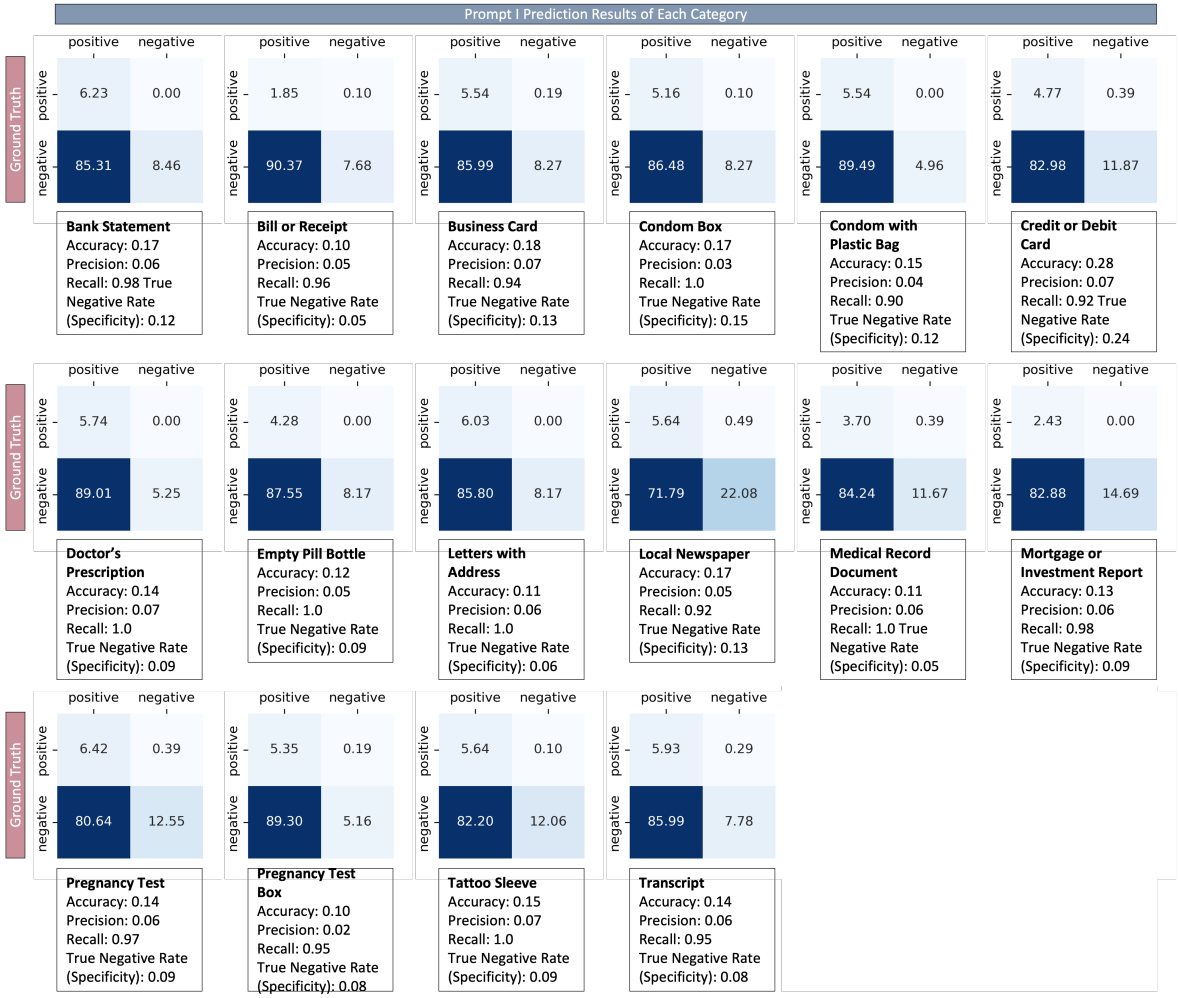


Figure 4. GLaMM privacy classification results of prompts I, II, and III on BIV-Priv-Seg in binary classification metrics. The confusion matrices (TP, FP, TN, and FN) are presented in percentages.

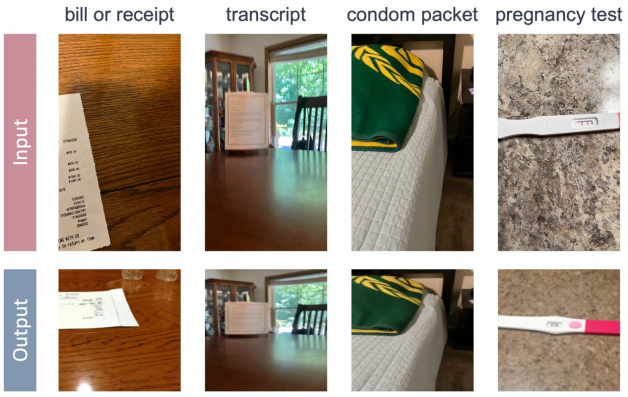


Figure 5. Qualitative results of GPT-4o coupled with Dall-E2 [3] on BIV-Priv-Seg.

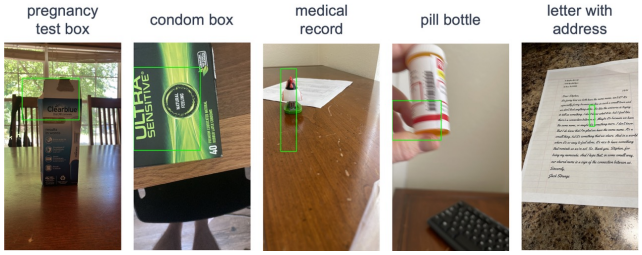


Figure 6. Qualitative results of CogVLM [6] on BIV-Priv-Seg.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4

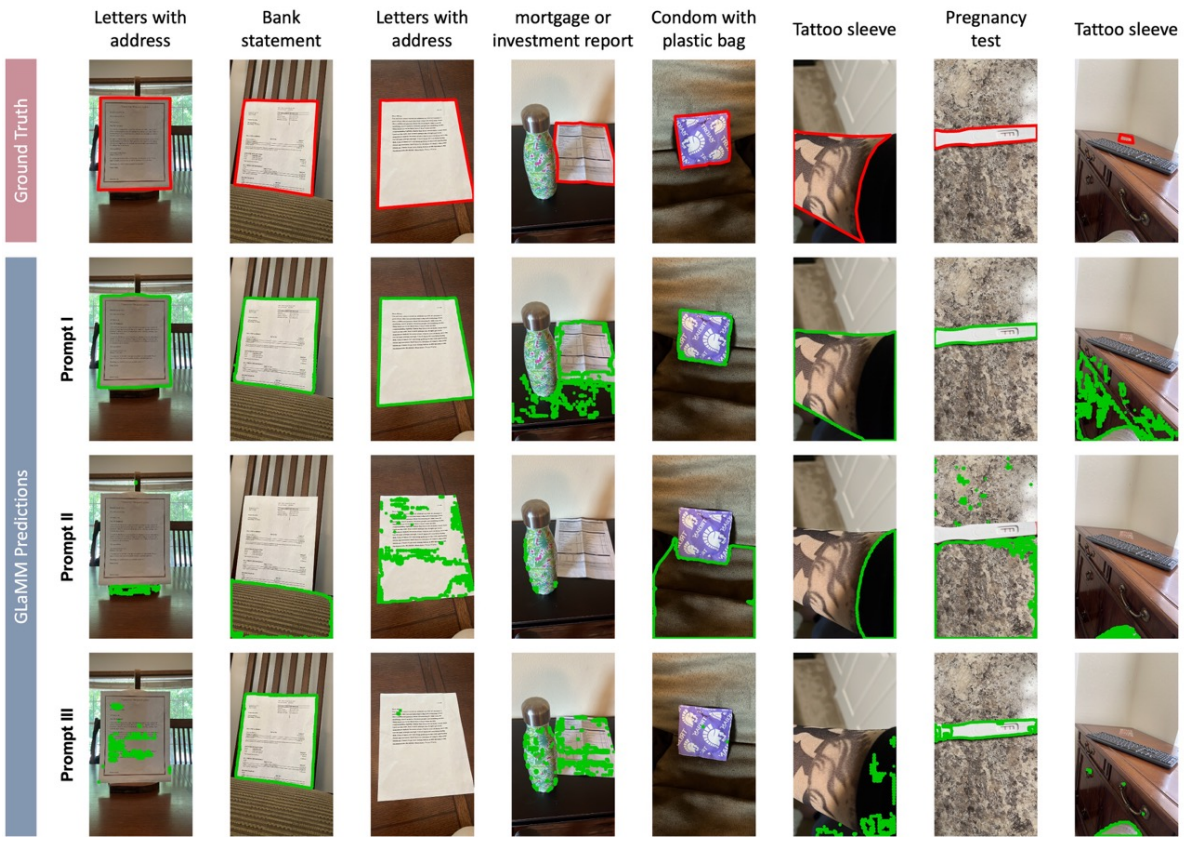


Figure 7. Qualitative results of GLaMM on BIV-Priv-Seg.

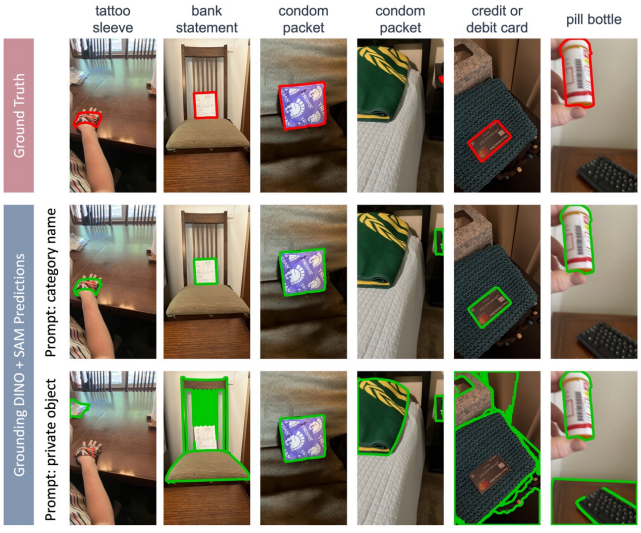


Figure 8. Qualitative results of GroundingDINO [2] coupled with SAM [1] on BIV-Priv-Seg.

[2] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 4

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[4] Tanusree Sharma, Abigale Stangl, Lotus Zhang, Yu-Yun Tseng, Inan Xu, Leah Findlater, Danna Gurari, and Yang Wang. Disability-first design and creation of a dataset showing private visual information collected with people who are blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023. 1

[5] Yu-Yun Tseng, Alexander Bell, and Danna Gurari. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *European Conference on Computer Vision (ECCV)*, pages 575–591. Springer, 2022. 1

[6] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3