Hierarchical Instance Tracking to Balance Privacy Preservation with Accessible Information - Supplementary Materials

Neelima Prasad[1], Jarek Reynolds[1], Neel Karsanbhai[1], Tanusree Sharma[2], Lotus Zhang[3], Abigale Stangl[4], Yang Wang[5], Leah Findlater[3], and Danna Gurari[1]

[1] University of Colorado Boulder, [2] Pennsylvania State University, [3] University of Washington, [4] Georgia Institute of Technology, [5] University of Illinois at Urbana-Champaign

This document supplements the main paper with additional information concerning:

A. Dataset Creation (supplements Section 3.1)
- Video Source
- Annotation Collection
- Ground Truth Generation

B. Dataset Analysis (supplements Section 3.2)
- Baseline Datasets for Comparison
- Dataset Composition
- Segmentation Properties

C. Model Benchmarking (supplements Section 4)
- Fine-Grained Analysis

A. Dataset Creation

Video Source. As noted in the main paper, two in-house annotators specified for each of the 552 videos the start and end frames when objects of interest where visible. We employed the Intersection Over Union (IoU) similarity score to gauge similarity among the annotator-flagged start and stop frames. For the intersection, we calculated the duration between the maximum value of the two annotated start times and the minimum value of the two annotated end times. For the union, we calculated the duration between the minimum value of the two annotated start times and the maximum value of the two annotated end times. We used an IoU threshold of 0.99 to determine whether the start and end frame annotations match.

Annotation Collection. We hired crowdworkers on Amazon Mechanical Turk to annotate our objects and parts with an annotation interface that we built. The interface collects a series of clicked points to create connected polygons on independent video frames. The interface supports annotating multiple polygons to capture when (1) there are multiple instances of a part (e.g., multiple account numbers) and (2) occlusions that break a part's appearance into multiple, disconnected pieces. Workers were given a comprehensive instruction set including instructions on how to segment each object class along with its parts.

To facilitate collection of high-quality annotations, we employed several quality control checks. We monitored ongoing quality by reviewing outliers regarding worker's frequency of indicating object and part non-presence, average time to complete a full annotation task, and the level of detail they provided in their segmentations (e.g., high prevalence of triangles). We conducted manual spot-checks at the conclusion of each phased task rollout.

Ground Truth Generation. We used redundant annotations to establish ground truth for objects.

We observed annotation agreement regarding the presence of an object for 96.5% of frames (present in 9,804 frames and absent in 971 frames), with 93% of the remaining 361 frames showing the object. Consequently, 91% (10,165) of the 11,165 annotated frames showed a target object. Of these, 98% (9980) were similar while 2% (185) had IoU scores less than 0.75 or lacked a redundant annotation necessary to calculate an IoU similarity score. For those lacking annotation agreement, the in-house annotators reviewed both annotations side by side and then chose one of the two annotations to keep for ground truth for 95% (175) of instances and resegmented the other 4% (10) where an object was missing or misidentified.

We observed annotation agreement that parts were not present for 43% (19,201) of 44,600 instances where crowdworkers were prompted about a part's presence. Of the parts deemed present, 67.8% (17,217) had high segmentation similarity and the remaining 32.2% (8,182) went through further manual review. An in-house annotator reviewed both part-level annotations and then selected the correct option when available or created a new segmentation when neither were suitable. Of the 8,182 part-level annotations, one part-level annotation was selected for 53.2% (4,357) instances and new segmentations were created for the rest.

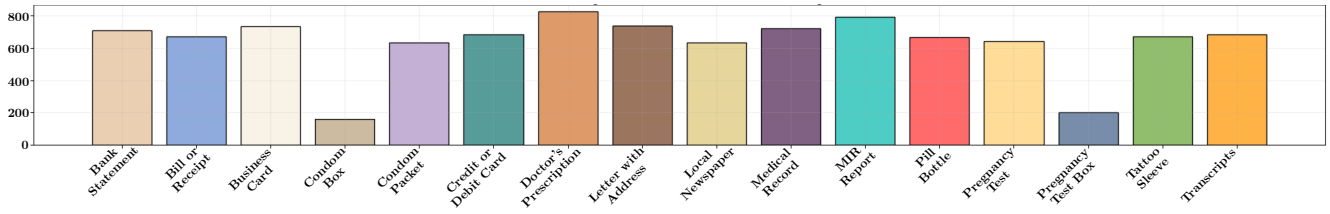

Figure 1. BIV-Priv-HIT object annotation frequency distribution of objects across all 10,165 object annotations

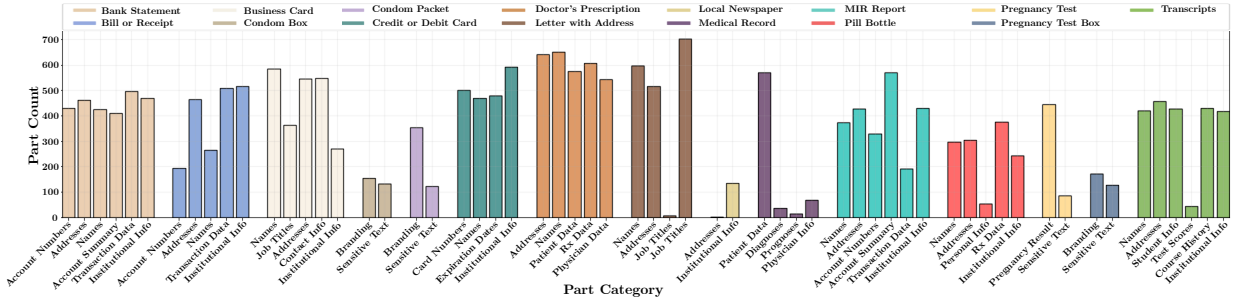

Figure 2. BIV-Priv-HIT part annotation frequency distribution of part categories across all object 10,165 annotations.

B. Dataset Analysis

Baseline Datasets for Comparison. Only one other dataset could feasibly support hierarchically tracking objects and parts, Meta's SA-V [2], since it provides both object and part masklets. However, it is non-trivial to determine the hierarchical object-part relations automatically. Specifically, inference is necessary because part and object masklets are treated the same, yet this is non-trivial to achieve for numerous reasons including that unrelated occluding entities can lead one to incorrectly deem an entity to be a part (e.g., a watch on a person's wrist).

Dataset Composition. The object category frequency distribution across the BIV-Priv-HIT dataset is shown in **Figure 1**. When observing the object category frequency distribution, the condom and pregnancy test boxes have the lowest object counts. This is because, in the original dataset, they are both categorized under the same label, 'condom' and 'pregnancy test.' In contrast, we observe that documents and similar objects feature the most part annotations per object. For example, bank statements, business cards, doctor's prescriptions, and similar objects feature the most part annotations per object. To ensure semantic labeling precision and increase granularity, we separated images featuring only the box versus the object and vice versa. In addition, we had comparable frame sampling across object labels, with every object label featuring between 650 and 800 human annotations per object.

We observe a similar trend in cumulative part annotations per object as in object annotations per object illustrated in **Figure 2**. We observe that condom boxes and pregnancy test boxes have the lowest number of part annotations per object because these labels also have the lowest number of object annotations. We also note that condom packets, medical records, and pregnancy tests have the highest occurrences of single-part annotations. This is because the condom packet and pregnancy test only have two parts, where one part (branding) is predominantly more visible than the other (sensitive text) in nearly all viewing scenarios. The medical record has four parts; however, the patient data part label features the highest visibility across diverse viewing scenarios, as it occupies the most significant amount of area relative to the object compared to its other parts (diagnoses, prognoses, and physician info). The most common frequency of part annotations across all objects is between 1 and 4 part annotations per object. Lastly, we found that Bank statements, MIR reports, and Transcripts are the only objects featuring 6 potential parts; however, only bank statements and MIR reports have instances where all six parts were annotated, and transcripts do not. Tattoo sleeves and local newspapers most often show no part annotations, which we attribute to a lack of parts for tattoo sleeves and no visibility of the private content for local newspapers.

Segmentation Properties. Statistics characterizing typical appearances of BIV-Priv-HIT's objects and parts are shown in **Figure 3**. In BIV-Priv-HIT, we observe that most objects feature boundary complexities between 0.65 and 0.75 (**Figure 3a**), while most parts feature boundary complexities between 0.35 and 0.60 (**Figure 3e**). The pregnancy test object features the most jagged and diverse boundary complexity, with 75% of its boundary complexities rang-

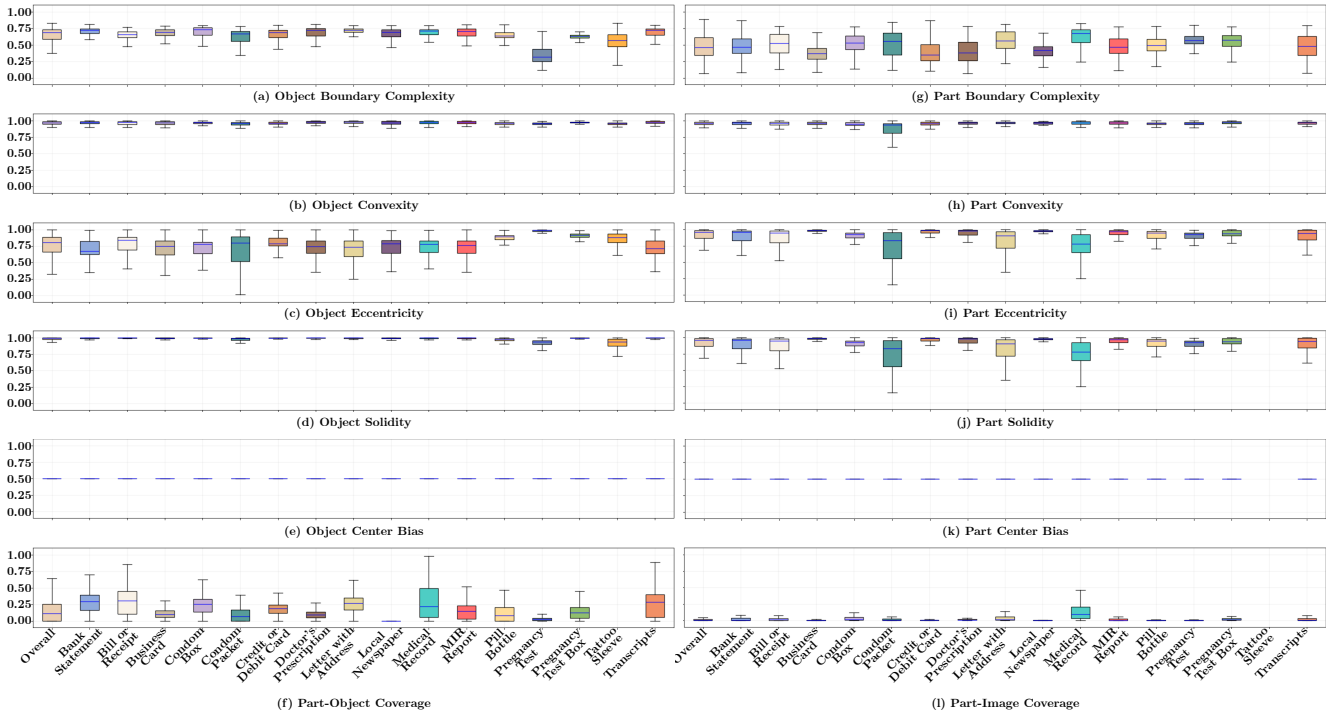

Figure 3. Boxplots showing the distribution of boundary complexity, convexity, and eccentricity at the object and part level. Part-object coverage and part-image coverage are also shown. The blue lines represent medians, bottoms and tops of each box represent the 25th and 75th percentile values respectively, and whiskers represent the most extreme data points not considered outliers.

ing from 0.25 to 0.42. We attribute this finding to pregnancy tests being the most geometrically complex objects out of all the object categories in the dataset. Moreover, the pregnancy test is the only object in the dataset that is not a square or rectangle and continuously presents complex boundaries regardless of the viewing angle. At the part level, the parts of the Business Card object feature the most jagged boundary complexities, with 75% of values ranging from 0.29 to 0.44. We attribute this finding to the inherent jagged edges caused by the occurrence of 'headings' and 'information.' For example, business cards typically feature a heading such as 'Job Title' or 'Email' followed by the information, which is the actual job title or email address. In many cases, the information is longer than the heading, so when annotating the part where we directed annotators to include the heading, the information naturally lends itself to creating multiple jagged edges due to including the heading and information in a single part annotation.

Regarding solidity (**Figure 3d**), nearly all objects and their respective parts are solid or 'filled' (solidity values closer to 1), illustrating that nearly all objects and their parts are their own convex hulls, and exhibit minimal indentations in their perimeters. At the object level, we observe that nearly all objects feature solidity values ranging from 0.96 to 1.0, meaning that nearly all of the object's pixels also fall within its convex hull. The two notable exceptions to this observation are the pregnancy test. We attribute this finding to the pregnancy test being the most geometrically complex among the objects in the dataset, with the object's structure featuring several concavities. Similarly, the tattoo sleeve follows the shape of the arm from the elbow down to the wrist, lending itself to an inherently indented perimeter shape. We see slight variations in other objects but attribute these variations to viewing angles, occlusions, and other artifacts that can potentially alter the object's relative convexity, for example, viewing a document nearly straight on as opposed to from the top-down.

We see a similar phenomenon at the part level (**Figure 3j**); however, the objects with the more diverse solidity at the part level are the condom packet and the pill bottle. In the case of all the objects, we observe a similar trend to the object level: nearly all parts are 'solid' with solidity values ranging from 0.95 to 1.0. Regarding the condom packet and the pill bottle, the exceptions to this trend, we attribute the increased convexity to the fact that these two objects feature a significant presence of text. In the case of the condom packet, the sensitive text is also placed among the branding, causing annotators to create more concavities in their annotations to segment sensitive text accurately. We see a similar trend in the pill bottle object due to parts such as addresses, personal information, and prescription data, all of which are shapes that require more significant concavi-

ties in their segmentation to accurately demarcate from the other parts.

Regarding center bias (**Figure 3e**), values close to 0.5 indicate a balanced distribution of objects within the frame, suggesting that objects are neither heavily centralized nor significantly off-center. The dataset's median center bias value is approximately 0.4997, and all objects feature a narrow center bias range between 0.49 and 0.5, indicating a precise and slight central tendency at the object level. At the part level (**Figure 3k**), we see the same median value of 0.4997, albeit the spreads and whiskers are slightly wider than the object level, indicating more variability in the positioning of parts within objects.

Regarding convexity (**Figure 3b**), we see similar trends to solidity at the object level, with convexity values ranging from 0.94 to 0.99. This finding suggests that the shapes are relatively smooth at the object level and lack significant indentations or concavities. We see almost an identical trend at the part level (**Figure 3h**), with convexity values typically ranging from 0.94 to 0.97. The only exception to this finding is the condom packet, which consists of two parts: sensitive text and branding. We find both of these parts to present many concavities due to the shapes required to segment sensitive text and branding accurately. For example, two of the condom packet brands found in the dataset are KY and Trojan; when segmenting the branding for these two condom packets, the KY logo and the Trojan helmet brand are shapes with many concavities and jagged edges. As a result, we observed that the parts of the condom packet had the broadest range of convexity values, ranging from 0.8 to 0.95.

Concerning eccentricity, at the object level (**Figure 3c**), we observe values exhibiting medians close to 0.8, a significant finding that indicates a generally high elongation in objects across categories. The interquartile range spans from approximately 0.62 to 0.98, further emphasizing the high median values. Moreover, we see a trend of whiskers extending from around 0.35 to 1.0, highlighting some eccentricity variation but maintaining a tendency towards higher values (more elongated). We see the condom packet's whiskers extend from 0.0 to 1.0, a finding we attribute primarily to the viewing angle because condom packets are only square-like when viewed top-down. In contrast, they can appear more elongated in nearly any other viewing scenario. We also observe high median values and tight spreads in the pill bottle, tattoo sleeve, and pregnancy test (median values ≥ 0.9), all of which are the most elongated objects in the dataset.

At the part level (**Figure 3i**), we see similar trends, albeit with higher median values (0.75 to 0.98) and less variance compared to the object level (whiskers primarily between 0.5 and 1.0). Again, we observe the condom packet elicits the most diverse eccentricity values, mainly due to the placement of sensitive text and the unique shape of their graphical brandings. We see a similar phenomenon in the letter with the address and medical record objects, which we attribute to the presence of text as these two objects consist of the most textually dense parts compared to other objects in the dataset. Overall, the eccentricity values at the part level generally show less variation than the object level, as the object's parts tend to have more defined and consistent shapes within their parent objects. We also provide solidity and center bias statistics at the object and part levels, detailed further in the supplementary materials.

Concerning part coverage, the relative area occupied by the region of interest, at the image level (**Figure 3l**) for nearly every object category, parts occupy less than 20% of the image with a majority of parts occupying less than 5% of the image. Again, we attribute the more significant inter-quartile range in the medical record object to the patient data, a part within medical records that can easily and often occupy more than half of the object.

At the object level (**Figure 3f**), we observe a similar phenomenon in that objects such as the bank statement, bill or receipt, medical record, and transcripts feature the largest interquartile ranges, a finding that we attribute to the relative sizes of the composite parts within these objects. For example, the transaction data part of a bank statement can and often does occupy most of the object compared to the account holder's name and address. Similarly, the grades part within the transcripts takes up most of the object's space instead of the student's name. In contrast, when examining the pregnancy test, we see a narrow interquartile range and a tight variance because the parts of the pregnancy test, such as the result and sensitive text, occupy very little space on the object itself.

We also report findings for adopting size thresholds introduced for the MSCOCO dataset [1], where 322 and 962 are thresholds determining whether an object is small, medium, or large. We find that in BIV-Priv-HIT's object annotations, 0.1% (13) of objects qualify as small, 2.9% (298) as medium, and 95% (9,854) as large. For part-level annotations, we find 6% (1,323) qualify as small, 41% (8,989) as medium, and 53% (11,725) as large.

C. Model Benchmarking

Despite the improved performance that comes from finetuning, our dataset still remains challenging for current state of the art models. **Figure 4a** shows that even a static object has so much variation in the predicted masks across frames. In **Figure 4b**, the model was unable to track all parts of the wrapper over time due to a shaky recording. **Figure 4d** not only has more than nine parts, but the object is under low lighting, which appears together to be challenging for the model, even at the object level.

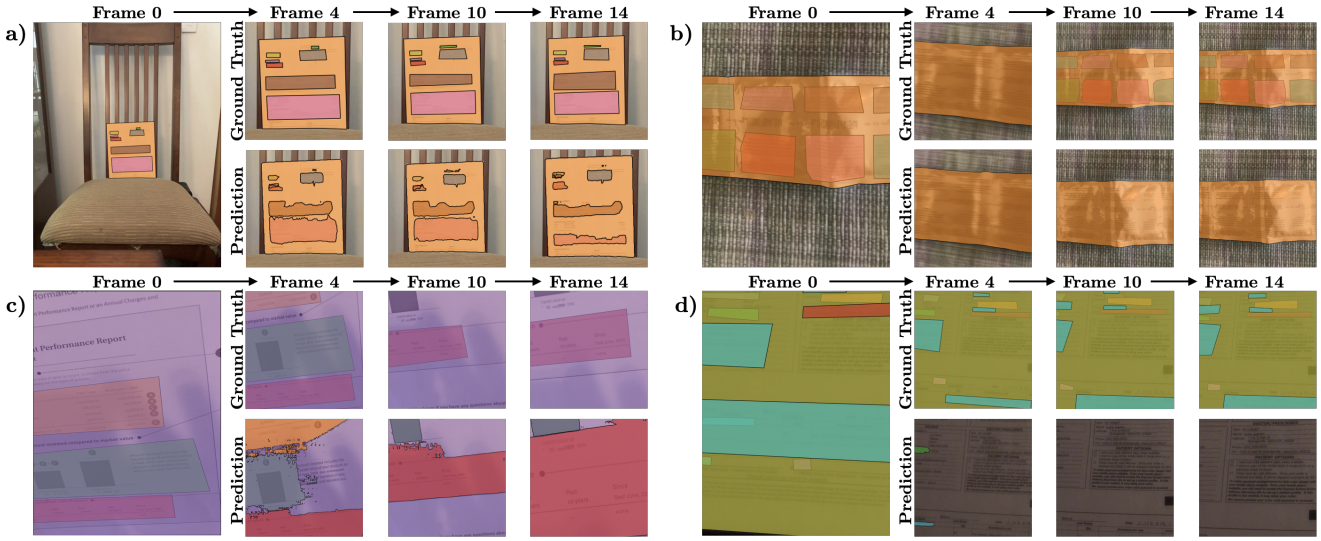

Figure 4. Examples of SAM-2's performance on frames collected from four video clips in our dataset. Shown is a full video frame with the ground truth mask (top) followed by cropped views of the ground truths and model predictions at subsequent frames in the video in order to make it easier to observe the model's performance on the region of interest.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[2] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2