Hierarchical Instance Tracking to Balance Privacy Preservation with Accessible Information

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

Abstract

We propose a novel task, hierarchical instance tracking, which entails tracking all instances of predefined categories of objects and parts, while maintaining their hierarchical relationships. We introduce the first benchmark dataset supporting this task, consisting of 2,765 unique entities that are tracked in 552 videos and belong to 40 categories (across objects and parts). Evaluation of seven variants of four models tailored to our novel task reveals the new dataset is challenging. Our dataset is available at this URL.

1. Introduction

Many people use camera-based services to stream videos showing their daily activities. For example, blind individuals regularly use them to learn about their visual surroundings [44, 45, 48, 60], including with Be My Eyes, Aira, and Envision AI. A growing number of sighted users are also using extended reality devices to enrich their daily viewing experiences, including with Meta's Orion glasses, VITURE's XR glasses, and Apple's Vision Pro. A key challenge for such video-based services is how to balance preserving privacy with retaining useful data.

Two types of scenarios underscore the tension between privacy preservation and data retention. *First*, is when a person shares their video feed with another person. This is common for blind people, who for example may want human confirmation regarding the required dosage for their prescribed medication in a particular pill bottle without revealing their name and address. AI could assist by either (1) obfuscating everything except the part of interest (e.g., dosage information) or (2) obfuscating only private categories (e.g., name and address). The *second* scenario is when a person shares video with a service provider that subsequently saves the data. It is common for blind people to share their private information with companies as a lesser evil to not learning about their visual surroundings [21], and in such cases obfuscating only the private categories (e.g., name and address) would preserve users' privacy while maintaining much of the utility of saved data for downstream purposes (e.g., training AI models). Importantly, for both these scenarios, an incorrect segmentation in even a single video frame would mean that information a user wants to conceal is revealed.

Addressing the need to balance preserving privacy with retaining useful data, we propose a novel task we call *hierarchical instance tracking*. It entails identifying and tracking all instances of predefined categories of *objects* and their *parts*, while maintaining their *hierarchical* relationships. This task unifies two problems historically examined independently: *video instance segmentation* (i.e.,tracking in videos all instances of predefined categories of *objects*) and *part segmentation* (i.e., locating in images all instances of predefined categories of *parts of objects*).

We introduce the first publicly-available dataset supporting our novel task, which includes annotations for tracking semantically-labeled objects and their parts using masklets (i.e., tracked segmentation masks). Notably, this is also the first publicly-available dataset to even semantically track just parts alone. We create the dataset by annotating 552 publicly-available videos taken by people with vision impairments of private content, which is called BIV-Priv [45]. For each video, we segmented and tracked every object and its parts that belong to 40 semantic categories, resulting in tracks for 2,765 entities with 537 objects and 2,228 parts. We call the resulting annotated dataset **BIV-Priv-HIT**, reflecting the task of **H**ierarchical **I**nstance **T**racking (**HIT**).

Next, we analyze how BIV-Priv-HIT compares to eight existing datasets for entity tracking and hierarchical segmentation. We show existing datasets *cannot support our novel task* because (1) none track parts *with* semantic labels and (2) none simultaneously track an object and its embedded parts, permitting the same pixel to belong to multiple

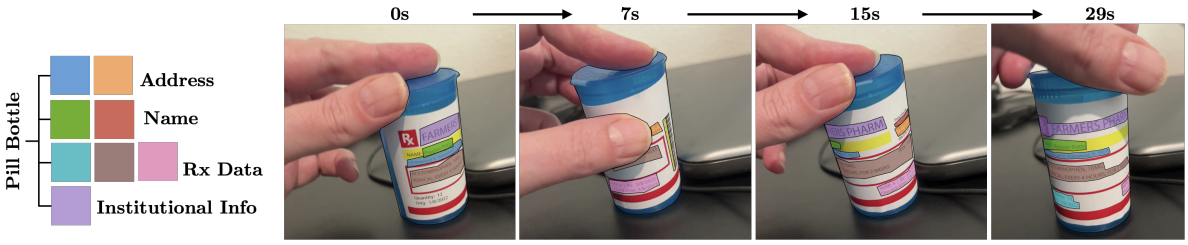

Figure 1. Example from our BIV-Priv-HIT dataset showing ground truth annotations we collected to track a pill bottle and its private parts. The legend on the left indicates the semantic labels and instance colors used to overlay tracked entities in the video frames.

semantic categories. We also show the new dataset *fills important gaps of existing datasets* and so supports developing more generalized algorithms. For instance, videos in BIV-Priv-HIT are orders of magnitude longer, ranging from approximately double to 11 times as long as videos in existing datasets. Additionally, segmentations in our dataset are unlike those in existing datasets including because (1) parts tend to contain text, (2) parts tend to have boundaries that are smoother and more elongated, at times even resembling a line segment, and (3) objects occupy much larger portions of video frames. This is exemplified in **Figure 1**.

Finally, we evaluate seven variants of four top-performing models for three related tasks—hierarchical image segmentation, video object segmentation, and video instance segmentation—on our dataset, after repurposing them for the task of hierarchical instance tracking. This effort includes introducing a new evaluation metric tailored to our tracking task. We found that all models perform poorly overall, especially for tracking parts but more generally for locating all small entities. In addition, the models are inefficient, as they require ad hoc workarounds involving multiple inference passes to perform our task. These findings underscore the value of our new dataset in supporting the AI community to tackle a new challenging problem.

We expect our dataset challenge will inspire new algorithmic designs for handling a greater diversity of real-world challenges within a single model. Success in this work could benefit other privacy-preserving applications, such as for individuals conversing with video streaming services (e.g., Zoom, WhatsApp), and robots navigating environments (e.g., emergency responders sent to crumbling/burning buildings). Success can also benefit other video-based applications by infusing finer-grained part-level understanding, including for robotics manipulation tasks, video editing, video retrieval, and pose estimation.

2. Related Work

Part and Hierarchical Segmentation Datasets. The recent successes of segmentation models at locating *objects* belonging to pre-defined categories has prompted a shift of focus for the community to instead segment more challenging *part-level categories*. This shift, which began in the mainstream computer vision community around 2017, has been inspired and enabled by new datasets that provide segmentations showing how objects are hierarchically decomposed into their nested parts [10, 15, 19, 23, 28, 34, 41, 46, 49, 51, 62–64]. Our work complements this literature by providing the first dataset containing semantic part segmentations as well as hierarchical segmentations for *tracking content in videos*. Additionally, parts in our dataset fill a gap of existing part-based datasets by exhibiting unique characteristics, including their tendency to contain text and so have smoother, more elongated boundaries.

Tracking Datasets. Two popular types of entity tracking datasets exist. *Video object segmentation* (VOS) datasets provide masks of entities tracked across all video frames (e.g., SA-V [42], DAVIS [40], YouTube-VOS [56]), while *video instance segmentation* (VIS) datasets also require labeling the category for each tracked entity (e.g., YouTube-VIS [54]). Most similar to our work is the SA-V [42] VOS dataset because it is the only other dataset that contains masklets (i.e., tracked masks) for *parts*. However, the SA-V dataset does not (1) provide semantic labels or (2) specify whether a masklet is for an object or a part (inferring this is non-trivial). Our work fills both these gaps.

Models for Segmentation Tracking and Hierarchical Segmentation. None of the models for tracking or hierarchical segmentation support our proposed task. For instance, video instance segmentation (VIS) models can be trained and applied for our target semantic object and part categories, but they cannot achieve this in a single inference pass since they do not permit the same pixel to belong to multiple semantic categories.[1] Similarly, video object segmentation (VOS) models [42] don't permit the same pixel to belong to multiple semantic categories and so would require

[1] An orthogonal line of research focuses on part-based tracking. While these methods also track "parts", their definition differs: in these approaches, parts are treated as appearance cues on objects (i.e., patches) that enhance the robustness of object tracking, rather than as semantic part categories with distinct identities [1, 6, 12, 14, 25, 37, 57, 58, 61].

multiple inference passes to support our task. Extending beyond VIS' limitations, VOS models also ignore semantics and require human annotation at the first appearance of each entity to track them. An alternative approach could be to apply hierarchical instance segmentation models [50] to every video frame to locate objects and parts, however such models ignore the fundamental concept of preserving "identities" over time and so would necessitate extra complexity to associate segmentation masks across video frames. Despite modern models' limitations, we benchmark them using ad hoc workarounds to highlight their potential value as a foundation for addressing our novel task. While experimental results reveal all types struggle, underscoring our dataset offers a challenging problem for the research community, VOS models offer the greatest promise.

Datasets Originating from Blind Individuals. This work also contributes to the movement in creating benchmark datasets where visual content originates from blind individuals. Most work focuses on images [2, 5, 7–9, 13, 20–22, 26, 29, 43, 47, 48, 59], with VizWiz [20] pioneering this direction in 2018, yet none provide hierarchical segmentations. Other efforts focus on videos [27, 31, 38, 52], yet none provide masklet annotations. Our work fills both gaps, contributing to the broader goal of designing more inclusive AI models that address the interests of blind people.

3. Hierarchical Instance Tracking Dataset

We now introduce BIV-Priv-HIT, the first dataset that supports hierarchical instance tracking (and part tracking).

3.1. Dataset Creation

Video Source. We leverage the 552 publicly-available videos from BIV-Priv [45], which were captured by 26 blind photographers of 16 private object categories [48] (e.g., credit cards). Importantly, none of the private content was pertinent to the photographer and instead originated from the datasets' authors with Institutional Review Board approval. Each photographer was instructed to capture an approximately 25 second clip for each type of private object twice, once with it positioned in the background to mimic *accidental* privacy disclosures and once in the foreground to mimic *intentional* privacy disclosures, emulating what was previously observed in authentic use cases [21].

Hierarchical Category Selection. We identified 24 *part* categories to associate with the 16 *object* categories established when curating the BIV-Priv videos [48]. We developed the part taxonomy to capture three tiers of common privacy concerns [18, 33, 39, 53]:

- *Personally Identifiable Information* (PII): information directly revealing an individual's identity, such as names, account numbers, and credit card numbers.
- *Quasi-personally Identifiable Information*: information that can indirectly reveal an individual's identity, such as addresses and job titles.
- *Sensitive Information*: not tied to a person's identity, but information a person might not want shared with others.

A list of all object categories and their associated privacy categories is shown in **Figure 2**.

Annotation Collection. For every video, we tracked every instance of each object and part category in our taxonomy using masklets. This involved a three-step process.

First, we constrained the annotation of masklets to only the clips of videos with target objects visible. To achieve this, three in-house annotators contributed by indicating the start and end frames where any target object category appeared (along with the detected category). For quality control, we had each video annotated by two people and then annotation differences were resolved via group discussions.

Next, we collected segmentations for every instance of the target objects and parts. We hired 25 trusted crowdworkers[2] from Amazon Mechanical Turk (AMT) to complete this using a home-grown interface which presents each frame and then has the annotator sequentially segment a specified object category followed by every visible instance of specified part categories. To accommodate occlusions fragmenting entities into multiple parts, we included a feature that enables creating multiple polygons when segmenting a single entity. We supported high annotation quality via on-boarding 'warm up' tasks, detailed instructions, live 'office hours' during annotation deployment periods for answering questions, and phased task rollouts to enable time for continuous inspection of submitted annotation results and worker feedback. Additionally, for each video frame, we collected annotations from two crowdworkers and then used their similarity to establish high-quality ground truth instance segmentations (as described in the supplementary materials). We chose to annotate every 40th video frame to balance annotating enough frames to be comparable in size to existing VOS datasets (see **Figure 2**) while skipping neighboring frames with very similar appearances. In total, 11,165 frames were annotated. This culminated in 10,165 object segmentation masks and 22,037 part segmentation masks, with 8.9% (i.e., 1,000) of video frames lacking segmentation masks. Cumulatively, the crowdworkers took 1,820 annotation hours (i.e., 45.5 40-hour work weeks) to complete the annotations.

Finally, in-house annotators associated the annotation masks belonging to the same entity across video frames and assigned unique instance IDs to each resulting masklet (for objects and parts). This was achieved using a home-grown tool, with quality ensured by having all resulting masklets verified by a second author. Cumulatively, these assocation

[2]We vetted the crowdworkers through their involvement in creating for us ∼40,000 object and part segmentations for other datasets.

	Ours	ADE20K [65]	PACO-EGO4D [41]	PACO-LVIS [41]	PartImageNet [23]	PASCAL-Part [10]
Tracking	✓	✗	✗	✗	✗	✗
# of Images	11.1K	27.6K	26.3K	57.6K	24K	19.7K
% of Images with Object Masks	91%	100%	90.9%	91.5%	100%	96.4%
% of Images with Part Masks	67.2%	45.7%	90.8%	91.5%	100%	96.4%
% of Objects with Part Masks	73.8%	13.6%	87.2%	76.4%	100%	80%

Table 1. Comparison of the composition of our dataset to existing hierarchical object-part segmentation datasets. While existing datasets hierarchically segment objects in images, BIV-Priv-HIT is the first to hierarchically segments objects across videos' frames.

annotations were completed in approximately 150 hours, which translates to an average of just over 3 minutes to associate all masks for each of the 2,765 tracked entities.

Dataset Splits. We divided the videos into training, validation, and testing splits using a 60% (327 videos), 15% (87 videos), and 25% (138 videos) split respectively. This resulted in the following number of annotated frames in the three splits: 6,690 in training, 1,680 in validation, and 2,795 in testing. When creating the splits, we ensured that the blind videographers who initially recorded the videos did not appear across multiple datasets splits to prevent models from overfitting to features of specific photographers.

3.2. Dataset Analysis

We now characterize BIV-Priv-HIT and how it compares to existing datasets.

Baseline Datasets for Comparison. We chose to compare our dataset to existing datasets that support the two distinct problems our proposed task unifies into the same framework: hierarchical instance segmentation in images and entity tracking in videos.

For *hierarchical segmentation* datasets, we chose those that similarly provide segmentations with semantic labels for objects and their nested parts. We chose the following recent and popular datasets: PACO-LVIS [41], PACO-EGO4D [41], ADE20K [65], PartImageNet [23], and PASCAL-Part [10].

We chose *entity tracking* datasets that similarly provide segmentation masks of entities throughout each video's frames to create *masklets*. This includes a popular video object segmentation dataset which provide masklets of objects *without* associated semantic labels—DAVIS [40]—and a popular video instance segmentation dataset which provides masklets of objects *with* associated semantic labels—YouTube-VIS [54, 55]. Also included is the recent SA-V [42] dataset which is the first dataset to provide part-level masklets, although without semantic labels or explicit flags indicating whether a tracked entity is an "object" or "part".

Dataset Composition. We first characterize and compare BIV-Priv-HIT's overall composition to existing datasets.

We report in **Table 1** a characterization of the *hierarchical segmentations* for all relevant datasets. As shown, BIV-Priv-HIT is unique because it is the only dataset to provide hierarchical segmentations for tracking entities in videos. All other datasets only provide segmentations for images. BIV-Priv-HIT also tends to have a smaller prevalence of annotated images showing at least one object of interest (i.e., 91%) and part of interest (67%), an especially important feature since tracked entities can disappear by leaving the field of view or becoming occluded.

	Ours	DAVIS [40]	YT-VIS [55]	SA-V [42]
Contains Semantics	✓	✗	✓	✗
Part Tracking	✓	✗	✗	✓*
Hierarchy Tracking	✓	✗	✗	✗
# of Videos	552	150	4,019	50.9K
Mean Length (sec)	27.9	2.4	5.31	14
# Annotated Frames	11,165	10,459	4,519	4.2M
# Instance Masks	32,202	27.1K	265.5K	35.5M
# Unique Instances	2,765	376	8,698	642.6K
Disappearance Rate	9%	16.1%	10.8%	42.5%

Table 2. Comparison of the composition of our dataset to existing entity tracking datasets. BIV-Priv-HIT is unique because it is the first to contain both semantic part annotations and hierarchical object-part instance segmentations as well as because it has longer video durations. (* flags that a dataset, in this case SA-V, lacks labels specifying whether any given mask is of an object or a part.)

We report in **Table 2** a characterization of all relevant *entity tracking* datasets. As shown, BIV-Priv-HIT is unique in three key ways. First, it is the only dataset to provide object and part segmentations with semantic labels. Second, it is the first to track hierarchically decomposed objects (i.e., of objects and their nested parts). Third, the typical video duration for BIV-Priv-HIT is orders of magnitude longer, with the mean video length of 27.9 seconds nearly double the mean length for SA-V [42] and over 11 times longer than the mean length for DAVIS [40]. Beyond these unique features, BIV-Priv-HIT is largely comparable to existing datasets. For instance, it lies in the middle of the pack with respect to size (i.e., number of included videos, annotated frames, and instance masks) and the number of unique entity instances. It also exhibits a similar prevalence of entity's disappearing in a video.

Next, we perform more fine-grained analysis of the prevalence of parts in BIV-Priv-HIT. We observe that different object categories can contain different amounts of parts (boxplots provided in the Supplementary Materials), ranging from many (e.g., bank statements contain up to 6 parts) to few (e.g., local newspapers contain at most 1 part) to none (e.g., tattoo sleeves). This underscores the value of our dataset in encouraging the design of models that can ac-

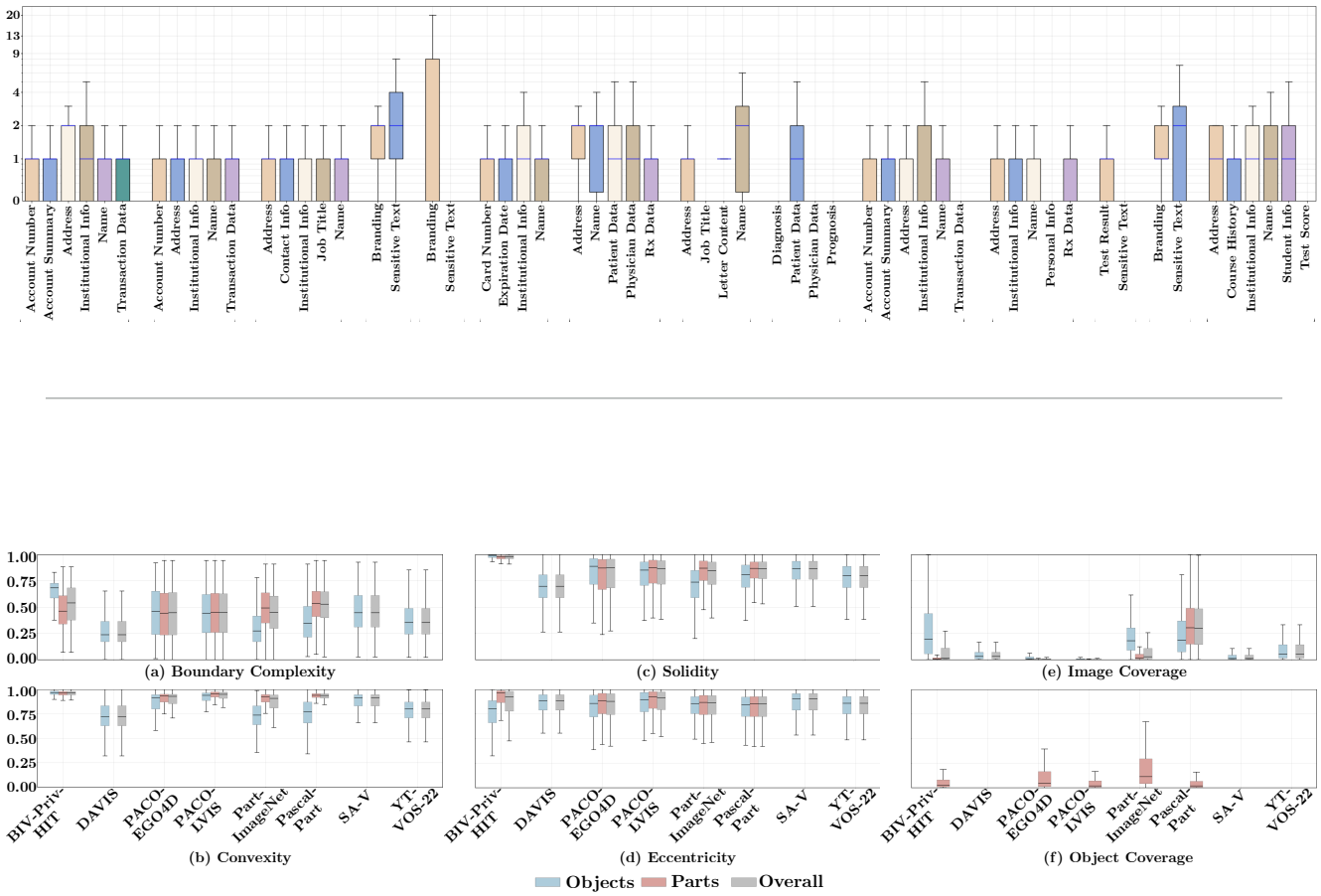

Figure 3. Boxplots characterizing segmentations in our dataset and seven other datasets with respect to six metrics, overall as well as with respect to only objects and parts independently. (Note: object coverage can only be computed for hierarchical segmentation datasets.)

count for object decompositions of different complexities. Additionally, within a specific object category, we observe variability regarding how many parts and which types are visible (**Figure 2**). Consequently, our dataset will encourage models to rely on visual evidence in each frame rather than biases from typical object-part associations.

Segmentation Properties. We also characterize the appearance of objects and parts and how they compare with segmented entities in related datasets[3]. To do so, we use the following metrics:

- **Boundary Complexity**: ratio of a entity's area to the length of its perimeter (*i.e.*, isoperimetric quotient). Values range from 0 (highly jagged boundary) to 1 (circular).
- **Convexity**: ratio of the entity's convex hull's perimeter to the perimeter of the entity. Values range from 0 (significant concavities) to 1 (perfectly convex).
- **Eccentricity**: ratio of the distance between a segmentation's foci and the length of its central axis. Values range from 0 (circular) to 1 (line segment).
- **Solidity**: ratio of the entity's area to the area of the entity's convex hull. Values range from 0 (fragmented or concave) to 1 (compact and solid).
- **Image Coverage**: ratio of the the image's pixels occupied by the entity (object or part). Values range from 0 (no image coverage) to 1 (complete image coverage).
- **Object Coverage**: ratio of the object's pixels occupied by all its nested parts. Values range from 0 (no object coverage) to 1 (complete object coverage).

Results are shown in **Figure 3**. When comparing BIV-Priv-HIT to all the entity tracking and hierarchical segmentation datasets, we observe segmentation masks in our dataset are distinct for five of the six metrics (not object coverage). We anticipate these distinctions, described below, will encourage the design of models that can generalize to a greater diversity of segmentation mask types.

With respect to the *entity boundary*, we observe that the objects in BIV-Priv-HIT exhibit the least complexity

[3]For datasets larger than BIV-Priv-HIT, we randomly sample 10,165 annotations for analysis to avoid excessive computational cost. For datasets with part annotations, we sample objects with at least one part annotation. For datasets smaller than BIV-Priv-HIT we use the entire dataset.

(**Figure 3a**), convexity (**Figure 3b**), and solidity (**Figure 3d**). This makes sense since our dataset focuses largely on human-made artifacts that typically have rectangular shapes, such as documents (i.e., pieces of paper), boxes, and pregnancy tests. Such human-made artifacts, unless damaged, lack concavities and jagged edges.

With respect to the *entity shape*, objects tend to be more circular than elongated (e.g., line segment) while parts tend to be more elongated than circular compared to existing datasets (**Figure 3c**). We attribute the latter observation to parts typically being textual information, which manifests as a line segment.

With respect to *entity size*, BIV-Priv-HIT's objects occupy the greatest diversity of sizes while parts occupy the least diversity of sizes compared to existing datasets (**Figure 3e**). We attribute the former finding to the fact that the objects of interest were intentionally positioned both in the background and foreground of images by the photographers. Moreover, objects can occupy larger portions of an image than observed in other datasets. This finding aligns with prior work's findings [7, 43, 47], which noted that people with vision impairments take close-up photographs of objects to better facilitate visual interpreters to recognize and so describe the visual content.

Semantic Properties. The 40 semantic categories in BIV-Priv-HIT share little overlap with all other datasets shown in **Tables 1** and **2**.[4] We attribute this to existing datasets' focus on content lacking private information, in accordance with best practices for dataset creation to remove such content. Our work, in contrast, *centers* on private categories.

Still, existing datasets can capture categories in our dataset with more abstract forms. For example, 'document' is a more general form of our 'bank statement' category. Additionally, PACO-EGO4D, PACO-LVIS, and PartImageNet all feature a category label for 'bottle' which shares a partial overlap with BIV-Priv-HIT's 'pill bottle'; however, the other datasets focus at the part-level categories on the composition and anatomy of the bottle (e.g., cap, neck, body, and label) while BIV-Priv-HIT focuses on the private contents of a pill bottle's label (e.g., address and prescription) [23, 41]. PASCAL-Part also features a label for 'card,' yet it makes no distinction as to what kind of card, such as credit card, playing card, business card, and so on [10]. Last, ADE20K shares the most partial overlap with BIV-Priv-HIT, with its category labels of bottle, card, bill, and document [64]. However, these objects do not contain part-level data and again represent more abstract, non-privacy-centric forms of the objects. While the category alignment with our dataset is limited, we suspect the few similarities across categories could facilitate models trained on the more abstract categories to generalize to the more specific categories encountered in our dataset.

[4]While SA-V could overlap with BIV-Priv-HIT due to its scale, SA-V excludes semantic labels preventing such comparison.

4. Evaluating Hierarchical Instance Tracking

Given the novelty of our task, we introduce a metric that for assessing how well models can preserve the hierarchical structure between an objects and its parts throughout tracking. Our key idea is to extend MOTA [4], a standard metric for multi-object tracking:

$$\text{MOTA} = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t}$$

where t is the frame index, FN are False Negatives, FP are False Positives, $IDSW$ are identity switches, and GT is the ground truth of one object.

We call our new metric MOTA-H, to reflect that it calculates tracking object's **H**ierarchical compositions such that parts of an object should remain associated with that object over time. Like MOTA, MOTA-H's score range from 1.0 for perfect tracking accuracy to negative infinity. Unlike MOTA, which determines detection matches between a prediction and ground truth using the overlap of bounding boxes, we instead use intersection over union (IoU) between segmentation masks, setting the IoU threshold to 0.5. We then change how identity switches are calculated to incorporate hierarchical relations as follows:

$$\textit{H-IDSW} = \begin{cases} 1, & \text{if predicted part ID changes} \\ 1, & \text{if predicted parent object changes} \\ 0, & \text{otherwise} \end{cases}$$

resulting in the following equation for MOTA-H, where all scores are only measured for the parts:

$$\text{MOTA-H} = 1 - \frac{\sum_t FN_t + FP_t + \textit{H-IDSW}_t}{\sum_t GT_t}$$

To also evaluate performance for tracking *objects*, we include a MOTA variant we call MOTA-OBJ, where the only change is computing detection matches between predicted and ground truth *objects* using IoU between segmentation masks (instead of the overlap between bounding boxes).

5. Model Benchmarking

We next benchmarked seven variants of four models. All experiments were run on NVIDIA's Tesla A100 GPUs.

Evaluation Metrics. We evaluate with respect to five metrics. Two are the metrics we introduced for our hierarchical instance tracking task: MOTA-H and MOTA-OBJ. The other three provide backward compatibility for analysis with video object segmentation (VOS), video instance

	Cheats	Inf./Vid. (Mean)	MOTA-H	MOTA-OBJ	J & F			AP			AR		
					Total	Objects	Parts	Total	Objects	Parts	Total	Objects	Parts
HIPIE-R50	✗	20.25	–	–	0.03	0.03	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HIPIE-ViT	✗	20.25	–	–	0.03	0.03	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2 Mask2Formers	✗	2.00	0.03	0.12	0.27	0.21	0.07	0.41	0.25	0.11	0.65	0.41	0.32
1 Mask2Former	✗	1.00	0.00	0.12	0.21	0.21	0.00	0.25	0.25	0.00	0.41	0.41	0.00
XMem++	✓	6.05	0.47	0.71	0.73	0.77	0.69	0.71	0.74	0.66	0.79	0.82	0.75
SAM-2	✓	6.05	0.39	0.54	0.58	0.73	0.53	0.58	0.82	0.53	0.59	0.76	0.55
SAM-2 Fine-tuned	✓	6.05	**0.72**	**0.76**	**0.78**	**0.90**	**0.77**	**0.76**	**0.90**	**0.74**	**0.83**	**0.93**	**0.82**

Table 3. Performance of benchmarking repurposed models for hierarchical segmentation (top), VIS (middle), VOS (bottom) on our dataset. (**Inf./Vid.** = Average number of inference passes per video)

segmentation (VIS), and hierarchical segmentation methods, but are all image-based (i.e., *ignore tracking*). First is $J\&F$ [40], the standard metric for VOS, which computes the mean between the Jaccard Index (J) (i.e., aka, intersection over union) and the boundary F-measure (F), the harmonic mean of precision and recall . The next two are average precision (AP) and average recall (AR), the standard metrics for VIS. Hierarchical segmentation papers [16, 30, 50] also use AP scores, with the key distinction from VIS methods that they report scores for both object and part categories (separately). The final three metrics all result in scores that range from 0 to 1, with larger scores (i.e., closer to 1) signifying better performance.

5.1. Hierarchical Image Segmentation

One relevant family of models for our task are those performing hierarchical instance segmentation. That is because they can provide the first critical step for tracking of locating all relevant objects and parts in each frame, and then leave it up to downstream association methods (e.g., Hungarian matching) to match segmentation masks across video frames to create masklets.

Model. We evaluate the top-performing hierarchical image segmentation model called Hierarchical Open-vocabulary Universal Image Segmentation (HIPIE) [50]. As noted in its name, the model is designed to support any vocabulary without further training. It outputs which categories are present where in a given image, when provided as input a list of all candidate categories that it should find. We feed the model as input our 40 object and part categories. We test two publicly-available variants, which rely on different backbones: ResNet-50 [24] and ViT-H [17].

Results. Results are shown in the first two rows of **Table 3**. Both variants failed completely, indicating no categories were present for nearly all the images where at least one relevant category was actually present. We attribute HIPIE's poor performance to poor generalization abilities, despite its claim to support an open vocabulary. Thus, we conclude current hierarchical image segmentation models are an inadequate foundation to extend for tracking, as they provide no detected instances for downstream associate methods to associate across frames.

5.2. Video Instance Segmentation

Another relevant family of models are those for video instance segmentation (VIS), which track specified semantic categories. While they cannot support our task in a single inference pass, since they do not permit the same pixel to belong to multiple semantic categories, a workaround is to instead develop two VIS models developed to support objects and parts separately and then leverage post-processing to associate parts with parent objects.

Model. We benchmark two variants of the popular, top-performing model called Mask2Former [11]. One variant results from fine-tuning two models for object and parts respectively, and then associating all masklets for parts to the masklet for the detected object, since a bias of our videos is they show only a single tracked object. The other variant results from fine-tuning a single Mask2Former to simultaneously support all object and part categories in our dataset.

Results. Results are shown in the second and third rows of **Table 3**. Overall, both methods perform poorly with respect to both the tracking-based and segmentation-based metrics. Moreover, both Mask2Formers performed worse for parts than objects, with the single Mask2Former trained to support both types neglecting parts entirely (i.e., scores of 0 for part-only metrics) in order to instead prioritize finding objects. We suspect the bias towards objects stem from a combination of this prioritization during initial training as well as that objects can be easier to locate, possibly due to their larger sizes and greater contrast of appearance relative to surrounding content.

Nonetheless, we observe a considerable performance boost from these models compared to the hierarchical instance segmentation model. We attribute this largely to Mask2Former being trained on in-domain data reflective of the test set. Moreover, it is promising to see that Mask2Former trained on only parts can succeed at times; e.g., AR of 0.32 and AP of 0.11. A valuable direction for future work is to explore if greater boosts can be secured

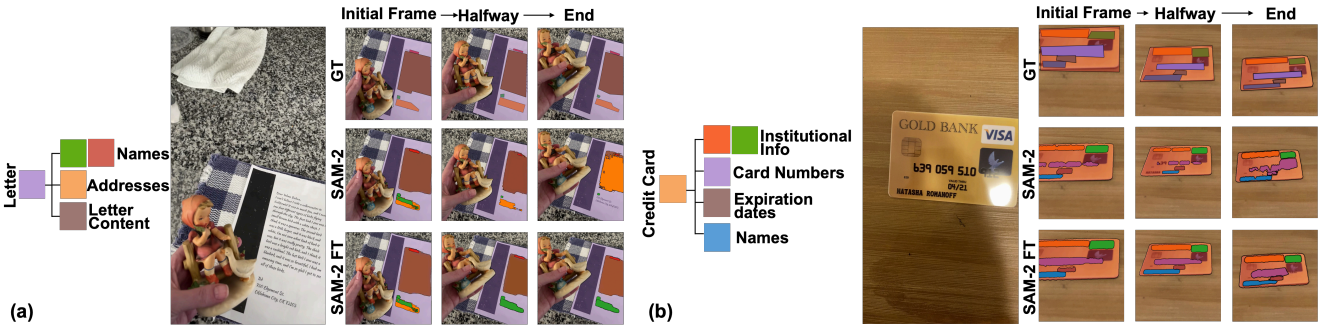

Figure 4. Initial video frames followed by cropped ground truths and predictions from SAM-2, as is and fine-tuned, in subsequent frames.

from more training data, or if fundamentally new architectures are needed.

5.3. Video Object Segmentation

Last, we evaluate video object segmentation (VOS) models. Like VIS models, they require multiple inference passes to recover objects and parts. Unlike VIS models, this is because VOS models only track one entity at a time, necessitating an inference pass for each object and part alongside post-processing to associate parts with objects. VOS models also require as input the target entity's segmentation in the first frame it appears.

Models. We evaluate two popular VOS models—SAM-2 [42] and XMem++ [3]—designed to track anything. Both are configured to *cheat*, receiving ground truth segmentation of each target entity in the first frame that each appears so they track entities only after knowing where to look when. This is necessary because automated localization from a model is not yet suitable (Section 5.1).

Results. Results in **Table 3** (rows 5-6) show these models outperform all other benchmarked models across metrics. This underscores a strong benefit of automating the cheating manual annotation step of locating target categories in images. This highlights a clear direction for future work: automating the manual annotation step we introduced, which currently provides the model with ground-truth part locations and artificially inflates performance.

Further analysis indicates that while VOS models often segment parts correctly in individual video frames, they struggle to consistently associate them across frames, leading to identity switching. This is reflect in high image-based segmentation scores (i.e., all exceed 0.5) alongside comparatively low tracking scores (i.e., MOTA-H below 0.5), suggesting that future work should also focus on improving mechanisms for associating part masks across frames.

Reinforcing earlier observations, a substantial performance gap remains between parts and objects. Suspecting this stems from a domain shift between SAM-2's original training data and our dataset, we fine-tuned SAM-2 on our training split.[5] The results in the final row of **Table 3** show a considerable improvement, particularly in MOTA-H, which nearly doubles from 0.39 to 0.72. We attribute this to SAM-2 learning a more effective appearance model for text-based parts, improving mask association across neighboring frames. Qualitatively, **Figure 4** shows the left sequence resolving identity switches between the "address" and "letter content" categories and the right sequence producing more uniform masks that exclude irrelevant regions (e.g. for the card number). While SAM-2 is designed to track *anything*, these results highlight that specialized datasets like BIV-Priv-HIT can substantially improve its tracking ability.

6. Conclusions

We introduce the novel task of hierarchical instance tracking along with the first dataset designed to support this task, called BIV-Priv-HIT. We also introduce an evaluation metric, MOTA-H to benchmark models' performance for the task. Our analysis reveals the unique characteristics of BIV-Priv-HIT compared to existing datasets. Benchmarking modern video object segmentation, video instance segmentation, and hierarchical image segmentation models demonstrate the dataset provides a challenging problem for the research community.

While this work advances hierarchical instance tracking, it has limitations. For example MOTA-H is based on MOTA, which can conflate different error types; other metrics (e.g., HOTA [36], TETA [32]) may provide a clearer assessment. The dataset also focuses on a limited set of part categories and domain, which may restrict generalization to other real-world scenarios. This work also carries ethical risks from misuse of models developed using the dataset, which future work should mitigate through developing responsible-use safeguards.

Acknowledgments. We thank Josh Myers-Dean for his assistance with setting up the models. This project was supported by National Science Foundation SaTC awards (#2148080, #2126314, and #2125925).

[5] XMem++ wasn't fine-tuned due to no public scripts or documentation.

References

[1] Osman Akin, Erkut Erdem, Aykut Erdem, and Krystian Mikolajczyk. Deformable part-based tracking by coupled global and local correlation filters. *Journal of Visual Communication and Image Representation*, 38:763–774, 2016. 2

[2] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16261–16270, 2023. 3

[3] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames, 2023. 8

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *Journal of Image and Video Processing*, 2008:246309:1–246309:10, 2008. 6

[5] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4271–4280, 2019. 3

[6] Elena Burceanu and Marius Leordeanu. Learning a robust society of tracking parts using co-occurrence constraints. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[7] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. 3, 6

[8] Chongyan Chen, Samreen Anjum, and Danna Gurari. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15315–15325, 2023.

[9] Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. Acknowledging focus ambiguity in visual questions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1228–1238, 2025. 3

[10] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. 2, 4, 6

[11] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation, 2021. 7

[12] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts, 2018. 2

[13] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3653, 2020. 3

[14] George De Ath and Richard M Everson. Part-based tracking by sampling. *arXiv preprint arXiv:1805.08511*, 2018. 2

[15] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 2

[16] Mingyu Ding, Yikang Shen, Lijie Fan, Zhenfang Chen, Zitian Chen, Ping Luo, Joshua B. Tenenbaum, and Chuang Gan. Visual dependency transformers: Dependency tree emerges from reversed attention, 2023. 7

[17] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[18] CSRC Content Editor. Pii - glossary: Csrc. 3

[19] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–785, 2018. 2

[20] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. 3

[21] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019. 1, 3

[22] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind, 2020. 3

[23] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2, 4, 6

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[25] Lianghua Huang, Bo Ma, Jianbing Shen, Hui He, Ling Shao, and Fatih Porikli. Visual tracking by sampling in part space. *IEEE Transactions on Image Processing*, 26(12):5800–5810, 2017. 2

[26] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Danna Gurari, Eunsol Choi, and Amy Pavel. Long-form answers to visual questions from blind and low vision people. In *Workshop on Demographic Diversity in Computer Vision@ CVPR 2025*, 2024. 3

[27] Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. A dataset for crucial object recognition in blind and low-vision individuals' navigation, 2024. 3

[28] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an

attribute localization dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 316–332. Springer, 2020. 2

[29] Jin-Hwa Kim, Soohyun Lim, Jaesun Park, and Hansu Cho. Korean localization of visual question answering for blind people. In *SK T-Brain-AI for Social Good Workshop at NeurIPS*, 2019. 3

[30] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity, 2023. 7

[31] Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. Exploring object status recognition for recipe progress tracking in non-visual cooking. *arXiv preprint arXiv:2507.03330*, 2025. 3

[32] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *European conference on computer vision*, pages 498–515. Springer, 2022. 8

[33] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020. 3

[34] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 2

[35] 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. 15

[36] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2): 548–578, 2021. 8

[37] Alan Lukežič, Luka Čehovin, and Matej Kristan. Deformable parts correlation filters for robust visual tracking, 2016. 2

[38] Daniela Massiceti, Lida Theodorou, Luisa Zintgraf, Matthew Tobias Harris, Simone Stumpf, Cecily Morrison, Edward Cutrell, and Katja Hofmann. Orbit: A real-world few-shot dataset for teachable object recognition collected from people who are blind or low vision, 2021. 3

[39] Weiping Pei, Yanina Likhtenshteyn, and Chuan Yue. A tale of two communities: Privacy of third party app users in crowdsourcing-the case of receipt transcription. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–43, 2023. 3

[40] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 4, 7

[41] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 2, 4, 6

[42] 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, 4, 8, 12

[43] Jarek Reynolds, Chandra Kanth Nagesh, and Danna Gurari. Salient object detection for images taken by people with vision impairments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8522–8531, 2024. 3, 6

[44] Laurens Samson, Nimrod Barazani, Sennay Ghebreab, and Yuki M Asano. Privacy-aware visual language models. *arXiv preprint arXiv:2405.17423*, 2024. 1

[45] 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, 3

[46] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5452–5462, 2019. 2

[47] 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*, pages 575–591. Springer, 2022. 3, 6

[48] Yu-Yun Tseng, Tanusree Sharma, Lotus Zhang, Abigale Stangl, Leah Findlater, Yang Wang, Danna Gurari Yu-Yun Tseng, and Danna Gurari. Biv-priv-seg: Locating private content in images taken by people with visual impairments. *arXiv preprint arXiv:2407.18243*, 2024. 1, 3

[49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[50] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7

[51] Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. Ov-parts: Towards open-vocabulary part segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[52] Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind, 2025. 3

[53] Anran Xu, Zhongyi Zhou, Kakeru Miyazaki, Ryo Yoshikawa, Simo Hosio, and Koji Yatani. Dipa2: An image dataset with cross-cultural privacy perception annotations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–30, 2024. 3

[54] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 4

[55] Linjie Yang, Yuchen Fan, and Ning Xu. The 4th large-scale video object segmentation challenge - video instance segmentation track, 2022. 4

[56] Linjie Yang, Yuchen Fan, and Ning Xu. The 4th large-scale video object segmentation challenge - video object segmentation track, 2022. 2

[57] Rui Yao, Qinfeng Shi, Chunhua Shen, Yanning Zhang, and Anton Van Den Hengel. Part-based visual tracking with online latent structural learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2363–2370, 2013. 2

[58] Rui Yao, Qinfeng Shi, Chunhua Shen, Yanning Zhang, and Anton van den Hengel. Part-based robust tracking using online latent structured learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6):1235–1248, 2017. 2

[59] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–31, 2020. 3

[60] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. Designing accessible obfuscation support for blind individuals' visual privacy management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024. 1

[61] Zhicheng Zhang, Shengzhe Liu, and Jufeng Yang. Multiple planar object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23460–23470, 2023. 2

[62] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 792–800, 2018. 2

[63] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1670–1678, 2018.

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 6

[65] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 4

Supplementary Materials

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.

B. Dataset Analysis

Baseline Datasets for Comparison. Only one other dataset could feasibly support hierarchically tracking objects and parts, Meta's SA-V [42], 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 5**. 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.

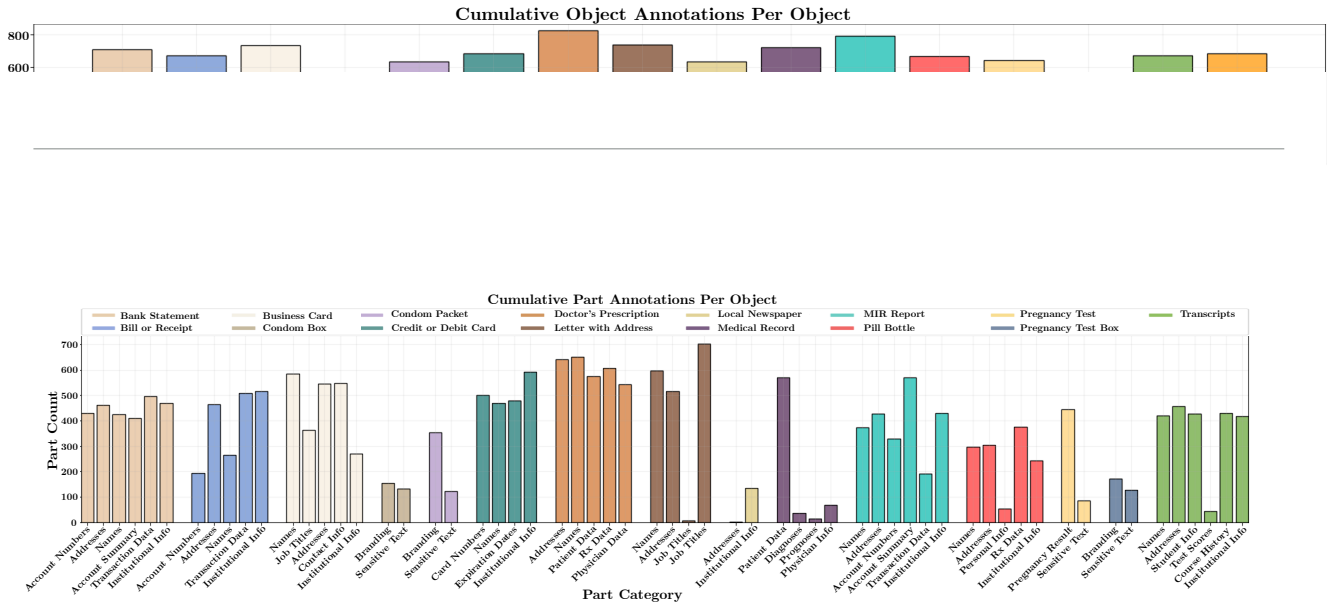

Figure 6. BIV-Priv-HIT part annotation frequency distribution of part categories across all object 10,165 annotations.

We observe a similar trend in cumulative part annotations per object as in object annotations per object illustrated in **Figure 6**. 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 7**. In BIV-Priv-HIT, we observe that most objects feature boundary complexities between 0.65 and 0.75 (**Figure 7a**), while most parts feature boundary complexities between 0.35 and 0.60 (**Figure 7e**). The pregnancy test object features the most jagged and diverse boundary complexity, with 75% of its boundary complexities ranging 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 7d**), 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

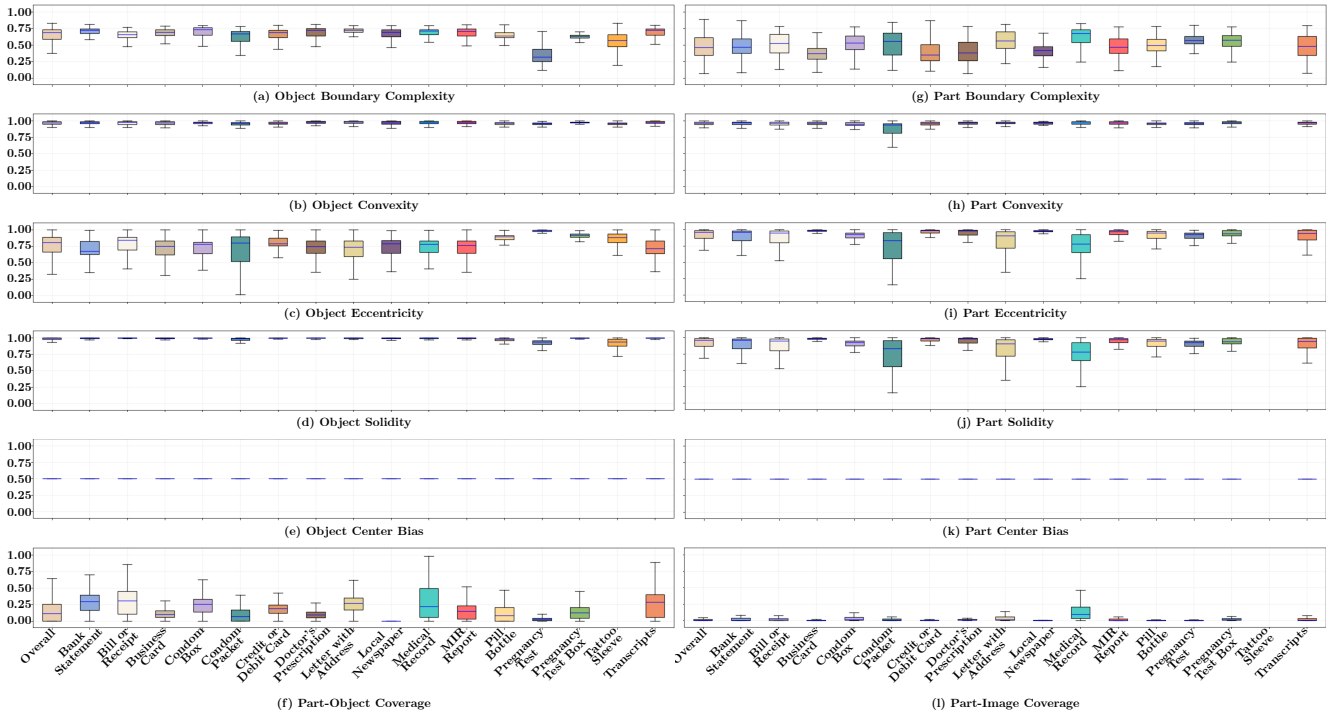

Figure 7. Boxplots showing the distribution of boundary complexity, convexity, and eccentricity at the object and part level. Part-object coverage and part-image coverage are also shown. The blue lines represent medians, bottoms and tops of each box represent the 25th and 75th percentile values respectively, and whiskers represent the most extreme data points not considered outliers.

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 7j**); 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 concavities in their segmentation to accurately demarcate from the other parts.

Regarding center bias (**Figure 7e**), 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 7k**), 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 7b**), 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 7h**), 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 7c**), 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 7i**), 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 7l**) 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 7f**), 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 [35], 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 8a** shows that even a static object has so much variation in the predicted masks across frames. In **Figure 8b**, the model was unable to track all parts of the wrapper over time due to a shaky recording. **Figure 8d** 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.

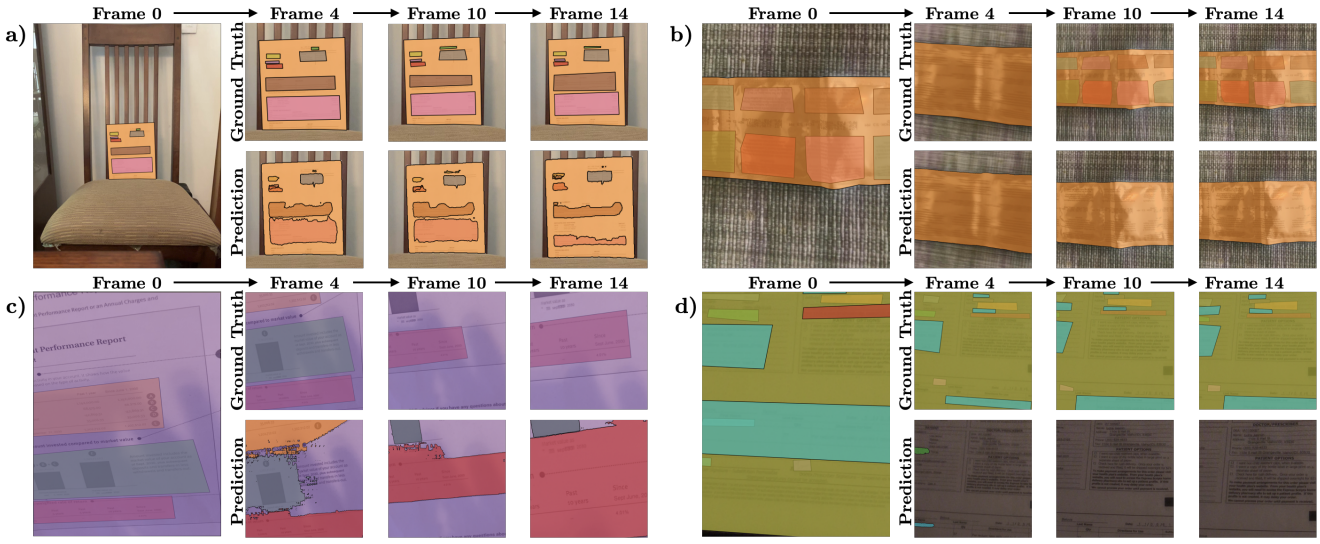

Figure 8. Examples of SAM-2's performance on frames collected from four video clips in our dataset. Shown is a full video frame with the ground truth mask (top) followed by cropped views of the ground truths and model predictions at subsequent frames in the video in order to make it easier to observe the model's performance on the region of interest.