BREPS: Bounding-Box Robustness Evaluation of Promptable Segmentation

Andrey Moskalenko[1,2,3,4*], **Danil Kuznetsov**[3*], **Irina Dudko**[3], **Anastasiia Iasakova**[3], **Nikita Boldyrev**[2], **Denis Shepelev**[2,3], **Andrei Spiridonov**[2], **Andrey Kuznetsov**[2], **Vlad Shakhuro**[1,2,3]

[1]Lomonosov Moscow State University
[2]FusionBrain Lab
[3]NUST MISIS
[4]IAI MSU
and.v.moskalenko@gmail.com, kuznetsov.danil@proton.me

Abstract

Promptable segmentation models such as SAM have established a powerful paradigm, enabling strong generalization to unseen objects and domains with minimal user input, including points, bounding boxes, and text prompts. Among these, bounding boxes stand out as particularly effective, often outperforming points while significantly reducing annotation costs. However, current training and evaluation protocols typically rely on synthetic prompts generated through simple heuristics, offering limited insight into real-world robustness. In this paper, we investigate the robustness of promptable segmentation models to natural variations in bounding box prompts. First, we conduct a controlled user study and collect thousands of real bounding box annotations. Our analysis reveals substantial variability in segmentation quality across users for the same model and instance, indicating that SAM-like models are highly sensitive to natural prompt noise. Then, since exhaustive testing of all possible user inputs is computationally prohibitive, we reformulate robustness evaluation as a white-box optimization problem over the bounding box prompt space. We introduce BREPS, a method for generating adversarial bounding boxes that minimize or maximize segmentation error while adhering to naturalness constraints. Finally, we benchmark state-of-the-art models across 10 datasets, spanning everyday scenes to medical imaging.

Code — https://github.com/emb-ai/BREPS.

1 Introduction

Promptable segmentation has rapidly transitioned from a niche task to a fundamental problem. Its rise coincides with the emergence of foundational Segment Anything (Kirillov et al. 2023) (SAM) model and successors. These models can output pixel-accurate masks of objects referred by a simple user prompt, e.g. point, bounding box, text, or a coarse mask. Nowadays promptable segmentation models are widely adopted to downstream applications ranging from photo and video editing to semi-automatic data labeling (CVAT (Sekachev et al. 2020)) and perception for robotics.

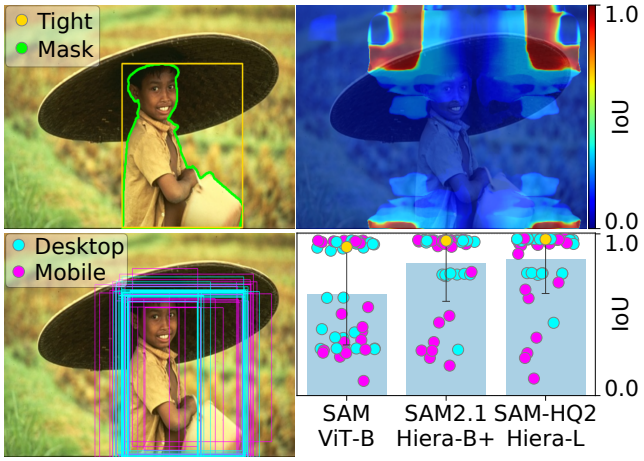

Figure 1: Top-left: image with ground-truth mask (green) and tight bbox (yellow). Bottom-left: real-users bboxes — desktop (cyan) and mobile (magenta). Top-right: IoU heatmap for SAM ViT-B; pixels shows IoU for a bbox anchored at that pixel and centered on the object. Bottom-right: IoU spread across 3 SOTA models for different users. We observed the large variability in IoU between individuals.

Among the available prompts, a bounding-box (bbox) is the most informative. Models that use bboxes usually output the best first-round masks (Ravi et al. 2024; Mazurowski et al. 2023). These masks may be further improved with corrective prompts, i.e. points. Almost all existing training and evaluation protocols use a simple bounding box sampling procedure. They use *tight bbox* (e.g. bbox which is obtained from the boundaries of the instance on ground-truth segmentation mask) with a small jitter. This procedure doesn't include any prior information about how people actually draw bboxes. However, these models will ultimately be used by real users, who may encounter inconsistent model quality.

To gain insights on how models perform on real-users prompts, we conducted a large-scale user study with the help of 2,500 annotators. Surprisingly, we found that — even though human boxes cluster closely, resulting mask quality (IoU) varies significantly between annotators (see Fig. 1).

*These authors contributed equally.

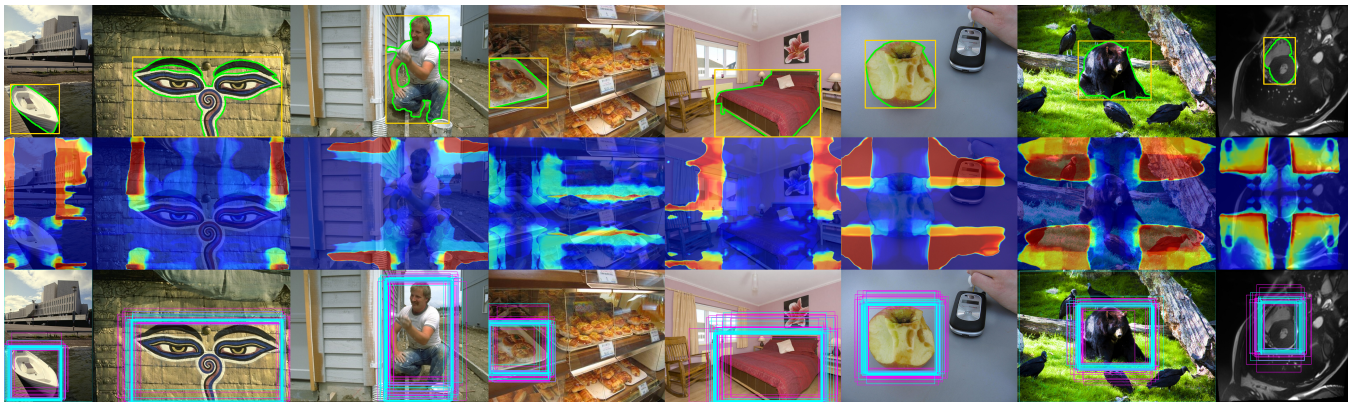

Figure 2: Top row: Berkeley, ADE20K, COCO, ACDC images with green mask and yellow tight bbox. Middle: SAM ViT-B IoU heatmaps—pixel color shows IoU (0 blue $\rightarrow$ 1 red) for a bbox cornered at that pixel and centred on the object. Bottom: user-drawn bboxes—desktop (cyan) vs mobile (magenta). Desktop prompts are tighter; user boxes vary and diverge from the tight bbox. Heatmaps reveal steep IoU drops from even 1-pixel shifts — examples provided in Supplementary. Zoom for details.

We believe this instability of the models is due to the overfitting to synthetic bounding boxes and causes sim-to-real gap when we employ these models in real-world applications.

Due to the limited availability of real-users prompts, we conducted a large-scale exhaustive search over valid bounding-box prompts for the most popular foundation models in both general-purpose and medical segmentation. We found that shifting the bounding box by just one pixel can change the quality significantly (see heatmaps in Fig. 1, Fig. 3). However, exhaustively searching over every plausible bounding box is computationally prohibitive. We therefore recast the problem as a white-box adversarial attack on the bounding box prompt space. A naive coordinate attack may collapse the box to a degenerate point; we prevent such trivial behavior using a regularizer derived from the empirical distributions obtained from our user study.

Finally, we distill these ideas into a new robustness metric and perform the first comprehensive evaluation of 15 promptable segmentation models across 10 public datasets from general to medical domain.

Overall, our main contributions are as follows:

- We conduct a pioneering controlled real-users study, collecting thousands of bounding box prompts across desktop and mobile settings. It reveals that users draw boxes that are far from the *tight bboxes* and that state-of-the-art promptable segmentation models exhibit significant variability in performance across users.
- We introduce BREPS attack, a white-box optimization method for generating adversarial bounding boxes, guided by differentiable naturalness constraints to ensure realistic prompt perturbations.
- We perform a large-scale evaluation of state-of-the-art segmentation models across 10 datasets, uncovering average performance gaps of 30% IoU under realistic prompt variation.

We believe that our methodology paves the way for promptable segmentation models that are more robust and higher-quality in real-world applications.

2 Related Work

2.1 Promptable Segmentation

Promptable segmentation task is a generalization of the interactive segmentation problem, where the user guides the model with positive and negative visual inputs, typically clicks, to segment accurately the desired object (Sofiiuk, Petrov, and Konushin 2022; Chen et al. 2022; Liu et al. 2023a; Zhou et al. 2023). Promptable segmentation models form a new paradigm in which a single, versatile network can be adapted to various tasks at inference time by user *prompts* — points, bounding boxes, text, or coarse masks.

The SAM (Kirillov et al. 2023) utilizes this idea by showing that a vision transformer (Dosovitskiy et al. 2021), trained once on a massive, automatically generated dataset, can generalize to novel objects, imaging modalities, and tasks with only a handful of user interactions. Since then, many SAM-like models have emerged, spanning general-domain (Ravi et al. 2024; Ke et al. 2023), specialized medical (Ma et al. 2024; Cheng et al. 2023), efficiency-oriented (Zhang et al. 2023a,b), robustness-enhanced (Chen et al. 2024; Rahman et al. 2024) models, and extensions that link grounding or text prompts to segmentation (Ren et al. 2024; Liu et al. 2023b).

Among the rich prompt vocabulary, bounding boxes consistently yield the highest first-shot mask quality, often matching or even surpassing several corrective points in mean IoU (Ravi et al. 2024; Mazurowski et al. 2023).

2.2 Adversarial Robustness

Current benchmarks emphasize generalization across images and classes, but rarely across the prompt space itself. For example, previous works mostly sample bboxes from simple distributions (e.g. the *tight bbox* with a small jitter), omitting the question of how real-users draw boxes and how sensitive models are to plausible variations. As a result, a model may achieve a strong average IoU under a single canonical prompt while exhibiting large performance variance under equally reasonable real-world alternatives.

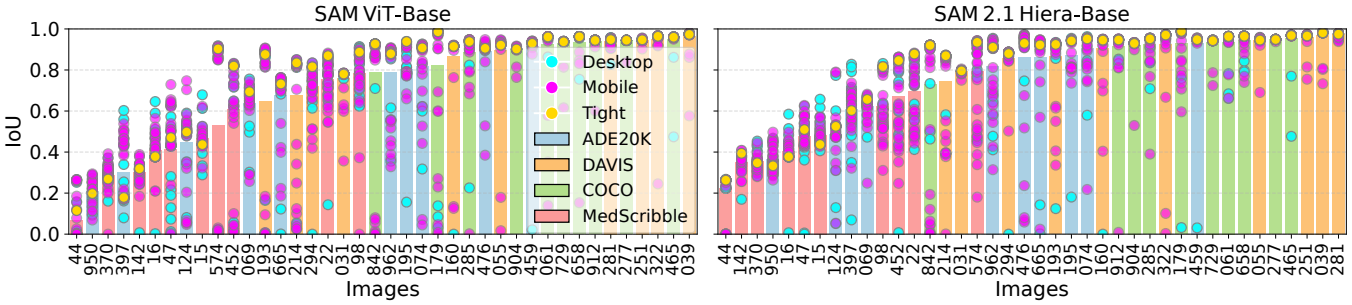

Figure 3: IoU spread for 10 random instances (3 general + 1 medical datasets). Points: desktop (cyan), mobile (magenta), tight bbox (yellow). Columns sorted by mean IoU over 50 users. SAM and SAM2.1 show large inter-user variance, while *tight bboxes* consistently over-estimate quality.

Conventional adversarial attacks (Goodfellow, Shlens, and Szegedy 2014) focus on perturbing the image. While promptable segmenters excel at generalizing across visual domains, some works have revealed that they are unstable to adversarial attacks (Wang, Zhao, and Petzold 2024). (Croce and Hein 2024) demonstrated that a geometry-invariant pattern pasted onto any input can cause SAM to output empty masks, regardless of the user prompt. DarkSAM (Zhou et al. 2024) uses additive pixel noise to force the model to predict the background for any prompt. BadSAM (Guan et al. 2024) implements a finetuning procedure such that a specific trigger prompt yields a preselected mask. RoBox-SAM (Huang et al. 2024), PP-SAM (Rahman et al. 2024) and ASAM (Li, Xiao, and Tang 2024) tuning the model with randomly perturbed prompts so it remains accurate under box or point jitter. Yet they assume small, synthetic perturbations and therefore fail to cover real-users behavior as in our study.

TETRIS (Moskalenko et al. 2024) probe the robustness of interactive-segmentation models to click prompts by using gradient-based optimization. Unfortunately, the resulting points are not realistic from a human-annotator standpoint — for example, they can land precisely on object boundaries, a behavior real-users rarely exhibit. We address this gap by introducing a regularization that explicitly constrains the search to *human-plausible* bounding box prompts.

The authors of RClicks (Antonov et al. 2024) train a clickability model on a corpus of genuine user clicks and leverage it to benchmark the accuracy of click-driven models. However, their evaluation is limited to a black-box setup with random sampling, and it entirely omits the widely used — and often higher-performing bounding box prompt modality.

2.3 Segmentation Prompts User Studies

Early efforts to understand how people annotate images at scale focused on crowdsourcing visual object labels. (Su, Deng, and Fei-Fei 2012) quantified how quickly non-experts could draw *tight bboxes*. Their analysis uncovered systematic biases — e.g. small objects are often missed — that later motivated tighter interaction loops.

Building on this, a line of work examined how much human effort each supervision modality actually costs. (Papadopoulos et al. 2017) introduced Extreme Clicking — four corner clicks instead of a full box — to cut annotation time $\sim$7 s per object without hurting IoU. Click'n'Cut (Carlier et al. 2014) and DAVIS interactive video segmentation tracks (Pont-Tuset et al. 2017) extended the paradigm to pixel-level masks, demonstrating that several iterative rounds of semi-supervised segmentation with user feedback can match the quality of fully-supervised segmentation.

Several studies have analyzed real-users behaviour in interactive segmentation. (Myers-Dean et al. 2024) recorded free interactions and found that circling was preferred by users in more than 70% of cases, while less than 15% for clicks. Bboxes can be extracted from circling and put into the promptable segmentation models. RClicks (Antonov et al. 2024) crowdsourced corrective clicks and trained a clickability model to emulate human click distributions.

Despite advances, bounding box prompts remain underexplored. Existing analyses either treat boxes as noise-free rectangles or perturb them with small Gaussian jitter (Rahman et al. 2024), failing to capture the plausible yet "unlucky" boxes we observe in practice. In contrast to prior studies, we provide the first systematic look at bbox realism, variability, and adversarial vulnerability, laying the groundwork for robustness-aware training and evaluation protocols.

3 Real-Users Study

To investigate user behaviour when drawing bounding boxes we conducted a large-scale crowdsourcing experiment with 2,500 participants.

3.1 Data Selection

For this study, we required datasets that provide both reference images and ground-truth segmentation masks. To guarantee coverage across domains and use cases, we mixed general-purpose and medical segmentation sets. We sampled 50 instances from each of the popular promptable general and medical datasets (total # of instances after the dash):

- GrabCut — 50 (Rother, Kolmogorov, and Blake 2004);
- Berkeley — 100 (Arbeláez et al. 2011);
- DAVIS — 345 (Perazzi et al. 2016), subset from (Sofiiuk, Petrov, and Konushin 2022);
- COCO-MVal — 800 (Lin et al. 2014), subset from (Sofiiuk, Petrov, and Konushin 2022);

Method	Backbone	GrabCut	Berkeley	DAVIS	COCO	TETRIS	PASCAL	ADE20K
MobileSAM	ViT−T	88.92±11.49	83.67±10.73	76.54±10.21	84.46±10.27	75.65±11.56	82.05± 9.48	71.20±13.83
SAM	ViT−B	86.20±16.87	81.32±14.31	80.23±13.78	80.83±15.98	70.99±16.13	82.41±13.44	62.17±15.73
	ViT−L	88.59±13.45	84.50±12.85	83.85±12.36	85.16±13.80	79.57±14.46	86.15±10.69	64.41±14.55
	ViT−H	89.11±13.48	83.87±12.38	83.92±12.28	85.14±12.63	79.45±14.82	86.22± 9.98	63.68±13.02
SAM−HQ	ViT−B	91.74± 9.00	88.29± 9.69	83.92± 9.73	86.56± 9.53	80.80±11.10	85.99± 8.13	68.37±14.63
	ViT−L	92.15± 8.74	88.09±10.56	86.12± 9.92	**87.33**±10.21	**83.20**±11.91	**87.97**± 8.26	**71.24**±14.68
	ViT−H	92.30± 9.01	**88.44**± 9.63	**86.28**± 9.96	86.90±10.28	83.16±11.99	87.81± 8.06	68.68±13.83
SAM 2.1	Hiera−T	90.09±12.97	85.31±12.78	83.65±11.73	84.36±12.34	75.87±16.33	86.79± 8.99	68.56±14.56
	Hiera−S	90.36±13.06	86.51±12.64	82.57±14.25	84.60±13.47	77.14±16.00	85.82±11.05	66.44±16.07
	Hiera−B$^+$	89.72±14.12	84.90±14.07	83.39±15.59	84.80±14.11	78.38±16.81	87.18±11.27	68.40±15.39
	Hiera−L	90.31±13.18	85.82±12.91	85.67±11.97	85.45±13.12	78.74±15.36	87.26±10.60	65.36±15.63
SAM−HQ 2	Hiera−L	**92.35**±11.22	88.19±12.67	86.25±11.76	86.59±13.31	81.71±15.95	87.47±10.55	69.67±17.20
RobustSAM	ViT−B	78.39± 9.63	74.65±10.75	60.36±10.98	75.74± 9.44	62.06± 9.87	73.10± 8.56	54.29±10.58
	ViT−L	64.75±11.17	45.16±10.03	27.97± 7.41	52.70±10.15	42.07± 9.48	56.08±10.11	32.82± 9.91
	ViT−H	64.82± 6.40	49.30± 5.60	27.08± 2.99	61.54± 5.88	53.13± 6.12	58.90± 5.54	52.40± 6.79

Table 1: Promptable segmentation models IoU performance on real-users bounding boxes on general segmentation datasets. Best results are in **bold**, the second best is underlined. The standard deviation was computed across 50 users and averaged over all images in the datasets. Please refer to the Supplementary for separated desktop/mobile devices results.

- TETRIS — 2,531 (Moskalenko et al. 2024);
- ADE20K — 707,868 (Zhou et al. 2017);
- PASCAL-VOC2012 — 19,694 (Everingham et al. 2010);
- ACDC — 100 (Bernard, Lalande et al. 2018);
- BUID — 780 (Al-Dhabyani et al. 2020);
- MedScribble — 56 (Wong et al. 2024), 3–5 available samples per multiple medical datasets in the split.

Overall, from the 10 datasets we selected 500 images. To minimize load time, every image shown to workers was down-scaled so that its longer side did not exceed 1024px.

3.2 Crowdsourcing Setup

The design of the user-study begins with the choice of what can be considered a bbox and the method of display. The difference is that, unlike the modality of a point, which is made by a regular click, a bbox can be drawn in several ways while obtaining the same bbox by coordinates:

- The classic drag-and-drop from the top-left to bottom-right corner as used by CVAT (Sekachev et al. 2020), Label Studio (Tkachenko et al. 2020), and LabelImg (Tzutalin 2015);
- Two-click schemes recording the two opposite corners (two-point mode in CVAT (Sekachev et al. 2020));
- Polygons or lassos that are automatically converted to bboxes (V7 Labs 2024), (Supervisely 2024);
- Extreme clicking, i.e. four extreme points on the object contour (Papadopoulos et al. 2017).

Since drag-and-drop bbox labeling is the most widespread in current annotation pipelines, we adopted it in our study.

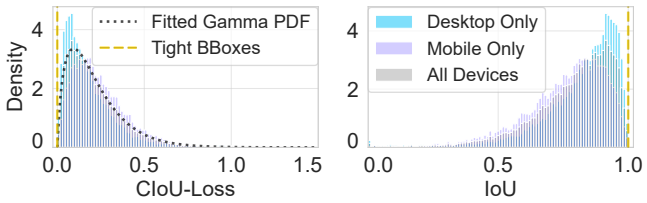

Figure 4: CIoU-Loss and IoU density plots for real-users drawn vs. *tight bboxes*. Black dashed shows fitted Gamma PDF. Mobile prompts (magenta) skew to lower overlap — we observe that on mobile devices, bboxes are more deviated from the *tight bboxes* than on desktop devices.

Display Bias Elimination Directly displaying the ground-truth mask would predispose participants and introduce anchoring bias (Draws et al. 2021). Our bbox drawing task, therefore, had to feel as natural as if users were selecting the object on their own.

Following the interface guidelines (Bauchwitz and Cummings 2025) and (Antonov et al. 2024), we implemented a three-stage protocol in our real-users study:

1. **Free viewing** — the raw image is shown for 3s; bbox drawing is disabled.
2. **Target memorization** — only the target region is visible for another 3s, the rest being grayed out; bbox drawing remains disabled.
3. **Annotation** — the mask disappears, the full image reappears, and the worker may draw exactly one bbox.
4. **Repetition** — if the assessor does not remember the target area mask, he can go through all three steps from the beginning by pressing the corresponding button.

Method	Backbone	ACDC	BUID	MedScribble
MobileSAM	ViT−T	79.29± 9.24	56.64±10.21	46.88± 8.54
SAM	ViT−B	83.43±10.10	61.19± 9.78	48.82±12.79
	ViT−L	**84.26**± 9.83	63.35± 9.06	<u>54.22</u>± 8.97
	ViT−H	83.09± 9.38	63.12± 9.26	53.73± 9.86
SAM−HQ	ViT−B	83.08± 8.52	58.39± 9.81	49.86±10.21
	ViT−L	<u>83.55</u>± 9.08	63.40± 8.48	53.10± 8.31
	ViT−H	82.83± 8.68	63.02± 8.90	53.35± 8.65
SAM 2.1	Hiera−T	80.21±10.87	60.15±12.56	**54.69**±10.20
	Hiera−S	78.87±10.16	63.20±10.32	54.10±11.66
	Hiera−B$^+$	80.34±11.11	61.27±12.24	51.85±11.00
	Hiera−L	79.11±10.52	<u>68.11</u>± 9.56	54.19±12.27
SAM−HQ 2	Hiera−L	79.54±10.06	**69.25**±10.19	53.13±11.68
RobustSAM	ViT−B	62.20±10.44	34.87± 7.20	46.01± 8.22
	ViT−L	43.09± 8.28	15.39± 4.46	27.63± 7.62
	ViT−H	46.81± 6.15	47.25± 5.34	36.77± 4.42

Table 2: Promptable segmentation models' performance on real-users bounding boxes on medical segmentation datasets. Best results are in **bold**, the second best is <u>underlined</u>. The standard deviation was computed across 50 users and averaged over all images in the datasets.

The annotation instruction for assessors:

- After opening the task, wait for the images to load within 30–60 seconds.
- After clicking the *Start viewing* button, you will be shown an image, then the area of interest on a gray background.
- Your task is to select **one rectangle** covering the area designated in the previous step.

The interface ran on both desktops and mobile devices. Each image was scaled to fill the available screen area; the bbox outline had a stroke width of 1% of the shorter image side. To compensate for finger thickness on touch screens on a near-border instances, a 5% from each side padding was added beyond the image borders. However, bbox drawings beyond borders results in coordinates clipping.

Each worker produced 10 bboxes for 10 different images. A hard time limit of 15 minutes was enforced, and the mean completion time was under 4 minutes. In all experiments, we use Toloka (Toloka AI 2025) crowdsourcing vendor.

3.3 Real-Users Behavior Analysis

In total, we collected 25,000 bounding boxes from 2,500 people (50 boxes per image), with half of the annotations drawn on desktop devices and half on mobile devices. In this section we analyze the collected user inputs, set bbox quality metrics and propose differentiable bbox realism regularization, which will be utilized in our BREPS pipeline.

Bounding-Box Quality Measuring We objectively measure the quality of observed bounding boxes w.r.t. the perfect *tight bbox*. Thus, we use the Intersection over Union (IoU) and the CIoU-Loss (Zheng et al. 2020) (Complete IoU), which is computed as follows:

$$\mathcal{L}_{\mathrm{CIoU}}\left(B, B^*\right) = 1 - \mathrm{IoU}\left(B, B^*\right) + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^*\right)}{c^2} + \alpha v$$

Where,

- $\mathrm{IoU}\left(B, B^*\right) = \frac{|B \cap B^*|}{|B \cup B^*|}$
- B and B^* denote the observed and ground-truth *tight bounding boxes*, obtained from segmentation masks.
- $\mathbf{b} = (x, y), \mathbf{b}^* = (x^*, y^*)$ are the centers of B and B^*.
- $\rho(\mathbf{b}, \mathbf{b}^*)$ is the Euclidean distance between the centers.
- c is the diagonal length of the smallest enclosing box covering both B and B^*.
- v measures the consistency of aspect ratios,

$$v = \frac{4}{\pi^2}\left(\arctan\tfrac{w^*}{h^*} - \arctan\tfrac{w}{h}\right)^2,$$

  where $(w, h), (w^*, h^*)$ are the B, B^* width and height.
- α is a positive trade-off parameter defined as

$$\alpha = \frac{v}{(1 - \mathrm{IoU}(B, B^*)) + v}.$$

Bounding-Box Quality Analysis The obtained distributions of the values of quality functionals are shown in Fig. As a ground-truth bbox, we took a *tight bbox* obtained from a ground-truth segmentation mask. We measured the performance separately for mobile and desktop users. We observed that users on phones are statistically significantly worse (U-test (Mann and Whitney 1947) with $p_{value} < 0.01$) in IoU and CIoU-Loss, which we attribute to the complexity of drawing on a small screen and the lower precision of a finger compared to a mouse cursor.

Measuring the Bounding-Box Realism We are interested in constructing a bounding-box realism regularizer for our optimization procedure. Since the CIoU-Loss exhibits smoother properties and allows gradients to flow even when the boxes do not overlap, we have chosen to build on it. We fitted several candidate distributions to a combined empirical sample (itself drawn from Beta and Gamma distributions). The best fit was achieved with a Gamma distribution ($k = 1.789, \theta = 0.121$) over the CIoU-Loss values between a bbox B and its *tight bbox* $B^\star$. Using the probability density function (PDF) of Gamma, the realism of any bbox B can be assessed directly through the log-likelihood ($\star$):

$$\log \mathrm{PDF}(X; k, \theta) = (k-1)\ln X - \tfrac{X}{\theta} - k\ln\theta - \ln\Gamma(k)$$
$$X = \mathcal{L}_{\mathrm{CIoU}}\big(B, B^*\big)$$

Obtained histograms and fitted PDF illustrated in Fig. 4.

3.4 Models Robustness on Real-Users

We compared 15 promptable segmentation models checkpoints using collected real-users prompts (MobileSAM (Zhang et al. 2023a), SAM (Kirillov et al. 2023), SAM-HQ, SAM-HQ2 (Ke et al. 2023), SAM2 (Ravi et al. 2024), RobustSAM (Chen et al. 2024). For these general segmentation models, we used all datasets from Sec. 3.1.

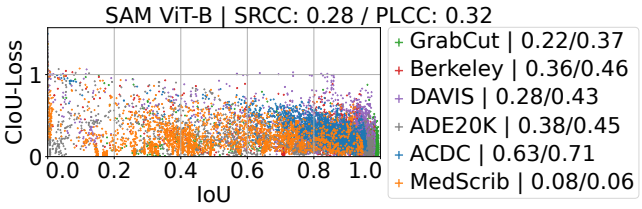

Figure 5: CIoU-Loss/IoU scatters for SAM ViT-B model; dataset labels show Spearman/Pearson correlations. We don't observe high correlations on 9 out of 10 datasets and associate the highest correlation of ACDC with the similarity and simplicity of instances in this dataset.

To compare models, we use the generally accepted IoU metric averaged across all instances in the dataset. Additionally, we present the standard deviation of model quality between different real-users. Results presented in Tab. 1, some examples provided in Fig. 3.

3.5 Bounding-Box Quality vs Model Performance

We explored whether the quality of a bounding box (relative to the ideal *tight box*) correlates with the quality of the resulting segmentation for that box. To do this, we compute Spearman and Pearson correlations for several models. A subset of the results for SAM is provided in Fig. 5.

We observed a weak correlation (≤ 0.4 SRCC on average on most datasets), indicating a more complex relationship between bbox perturbations and the resulting segmentation quality. This confirms the need for an assessment of the stability of promptable models and not just an assessment of the quality of the prompts at the input. Please refer to the Supplementary for more detailed analysis of collected bboxes.

4 Proposed BREPS Attack

4.1 Exhaustive Search

Since limited real-user and sampled prompts cannot cover every possible configuration of bounding boxes, we perform a computationally expensive exhaustive search on several instances from each dataset.

However, if we fully parameterize a bounding box by its four parameters (x_1, y_1, x_2, y_2) or (x, y, h, w), an image of size 1024×1024 (the model's standard input resolution) would require on the order of 1024^4 forward passes — clearly infeasible. To make the search somehow tractable, we restrict attention to boxes whose center coincides with the object's center, so we only vary the pair (h, w). Even this reduced search amounts to roughly one million inferences — still costly, but manageable for some instances.

For every bounding box, we obtain the model's prediction and compute the IoU between the prediction and the ground-truth mask. Next, for each pixel we plot the IoU obtained by the bounding box whose corner lies at that pixel. Since there are four such corner points, the resulting visualization (heatmaps for SAM (Kirillov et al. 2023) in Fig. 1, Fig. 2) is symmetric around the object's center.

We observed that the IoU heatmaps contain plateaus of high quality that abruptly drop off at certain boundaries.

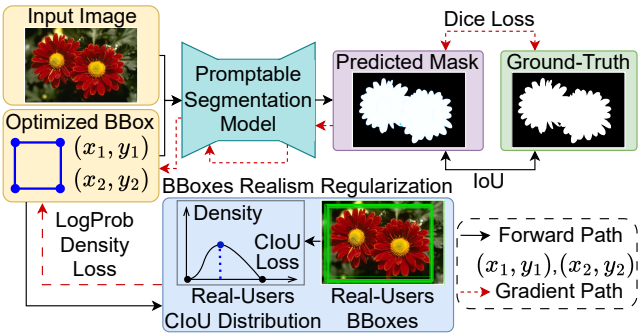

Figure 6: BREPS optimization pipeline.

Moreover, even within the high-IoU (red) regions, there are 'holes' where the quality is significantly lower. In other words, there exist bounding boxes that differ by only one pixel in their coordinates yet yield significantly different segmentation quality. In an ideal scenario, the model's output quality would be consistent for all reasonable user-drawn prompts (i.e., the heatmap would be a uniformly red area). This consistency would ensure user satisfaction in practical downstream applications.

4.2 BREPS Evaluation Protocol

Exhaustive search over all possible bounding boxes is intractable for large-scale datasets. We therefore reformulate the problem as a white-box adversarial attack in the space of bbox prompts. We observed that promptable segmentation models are differentiable with respect to their input prompts. Thus, we keep the image and the model weights frozen and perform gradient descent directly in the space of bboxes coordinates. The optimization pipeline is illustrated in Fig. 6. The gradient of the loss function passes through the attacked model and, adjusted with the realism regularizer, shifts the box coordinates accordingly.

We used the same datasets as in Sec. 3.1. Since we are no longer limited by the cost of human labeling, we included all instances from GrabCut, Berkeley, DAVIS, COCO-MVal, TETRIS, ACDC, and MedScribble, sampled 1,000 instances from ADE20K and PASCAL-VOC2012, and 503 instances from BUID (benign and malignant).

Realism constraint. Naively optimizing the segmentation loss often produces boxes that miss the object entirely, trivially lowering models' performance. To prevent such degenerate solutions, we introduce a realism term based on the Gamma density fitted in Sec. 3.3.

The overall objective is difference of DICE loss (Dice 1945) and λ-scaled log-probability density (Sec. 3.3, $\star$) of real-users CIoU-Loss distribution. For the IoU-maximizing attack, we negate the DICE term. We set $\lambda = 0.1$, choosing the Pareto-optimal trade-off between IoU degradation and log-probability realism (refer to the Supplementary).

Optimization details. First, we initialize the optimized box as the *tight bbox* in (x_1, y_1, x_2, y_2) parametrization. Then, we run 50 steps of the Adam optimizer to obtain realistic yet challenging boxes, taking under 2 seconds on an

Method	Backbone	COCO-MVal				PASCAL-VOC2012				ADE20K				BUID			
		Tight	Min	Max	Δ	Tight	Min	Max	Δ	Tight	Min	Max	Δ	Tight	Min	Max	Δ
MobileSAM	ViT-T	86.88	71.50	89.40	17.90	85.81	70.00	87.81	17.81	79.62	**49.66**	83.68	34.02	73.14	54.27	77.95	23.68
SAM	ViT-B	86.95	67.73	89.76	22.03	86.13	60.54	88.43	27.89	**79.93**	42.31	84.26	41.95	77.55	55.87	81.07	25.20
	ViT-L	87.63	71.06	89.36	18.30	88.21	71.26	89.40	18.14	79.43	39.20	83.48	44.28	75.39	**58.99**	78.73	19.74
	ViT-H	87.48	74.72	89.03	14.31	**88.46**	74.85	89.56	14.70	77.97	49.24	82.27	**33.03**	75.28	58.86	79.05	20.19
SAM-HQ	ViT-B	87.66	70.67	90.36	19.69	86.06	73.15	88.66	15.51	76.79	32.70	84.21	51.51	75.95	55.53	80.21	24.67
	ViT-L	87.77	74.68	89.90	15.22	87.92	77.46	89.63	12.17	79.73	33.11	85.25	52.15	75.28	58.52	79.52	21.00
	ViT-H	87.67	**76.73**	89.52	**12.79**	88.14	**79.89**	89.71	**9.82**	78.77	40.73	84.90	44.17	75.12	57.49	79.36	21.87
SAM 2.1	Hiera-T	86.87	55.80	89.76	33.96	87.21	46.36	88.55	42.19	77.69	22.72	82.95	60.23	77.63	43.90	82.23	38.34
	Hiera-S	87.09	52.55	89.60	37.05	87.51	46.88	88.70	41.82	75.34	15.46	80.34	64.89	76.53	45.74	80.64	34.89
	Hiera-B+	88.31	57.52	90.69	33.18	87.79	54.50	89.21	34.71	79.01	22.32	84.27	61.95	80.47	47.01	83.90	36.89
	Hiera-L	87.80	59.77	90.08	30.30	88.32	56.18	89.68	33.50	75.87	23.39	81.73	58.34	80.22	49.04	83.19	34.15
SAM-HQ 2	Hiera-L	**89.46**	66.60	**91.70**	25.10	88.44	71.64	**90.34**	18.69	79.60	23.12	**85.79**	62.66	**82.71**	51.95	**86.10**	34.15
RobustSAM	ViT-B	77.66	31.52	82.57	51.05	76.73	31.46	79.74	48.28	72.44	10.40	78.13	67.73	67.50	25.75	71.33	45.58
	ViT-L	52.06	4.22	57.92	53.70	68.40	5.05	73.92	68.86	53.85	1.86	61.31	59.45	29.16	2.29	34.77	32.48
	ViT-H	34.15	23.37	37.49	14.12	49.68	34.05	53.99	19.95	54.39	23.08	60.89	37.81	29.35	19.71	33.23	**13.52**

Table 3: BREPS attack results on state-of-the-art promptable segmentation models. Results provided for three general segmentation datasets and one medical one, the remaining are provided in the Supplementary due to limited space. Best results are in **bold**, the second best is underlined.

NVIDIA Tesla A100 for the SAM ViT-B model. Additionally, we apply a *clip* operation to ensure the bbox coordinates are inside the image and handle the edges in order (e.g. $x_1 < x_2$). Ablation on the number of steps is provided in the Supplementary. Since the optimizer operates in prompt space, whose scale varies between models (owing to different input resolutions), we linearly rescale the learning rate ($lr = 9$) with respect to the 1024×1024 input size of SAM.

Evaluation metrics. We adapt click-based metrics from (Moskalenko et al. 2024) for robustness evaluation on bounding boxes:

- **IoU-Tight@BBox**: IoU with the *tight bbox* prompt;
- **IoU-Min/Max@BBox**: IoU on the worst/best-case bbox found by the quality-decreasing/increasing attack;
- **IoU-Δ@BBox**: robustness metric defined as the difference between Max and Min attacks.

4.3 Discussion

Evaluation results on several datasets are shown in Tab. 3. More datasets are provided in the Supplementary. Based on the results, we made several conclusions:

- Following the results of our real-users study (Fig. 1, Fig. 2, Fig. 3, Tab. 1, Tab. 2) and robustness evaluation (Tab. 3), we can conclude that state-of-the-art interactive segmentation models are **extremely sensitive to the bounding box prompt fluctuations**.
- In the real-users study (Fig. 3, Tab. 1, Tab. 2) we also observed a strong spread in quality between user bboxes; on average, this **spread is about 15% of the IoU quality**.
- During the exhaustive search, we found (Fig. 2) that there exist **neighboring bbox positions that differ significantly in quality**. We note that the revealed problem is **relevant for both general and medical domains**.
- We observed a significant drop in quality when optimizing for minimization. On average, the **quality drops by 30% IoU relative to *tight bbox***, while according to our optimization strategy, bboxes still lie in the distribution of human-probable bboxes. This indicates potential quality drops when using such models on real people and highlights their overfitting to *tight bounding boxes*.
- Moreover, we observed that ***tight bboxes* turned out to be suboptimal in terms of maximum possible quality**, meanwhile they **overestimate real-world performance**. As a result of our optimization for IoU maximization, the quality of the models can be **increased by around 3% IoU** on average across all 10 datasets.

5 Conclusion

In this work, we explore the robustness of promptable segmentation models. Firstly, we gathered 25,000 real bboxes from 2,500 annotators using crowdsourcing. We also provided statistics and described the distribution of the obtained real-users bboxes. On these prompts, we evaluate state-of-the-art models and observe a large quality spread between users. We observed this effect in both general and medical segmentation domains. Then we performed an exhaustive search sweeping millions of plausible bboxes per instance, and revealed IoU gaps even for neighboring pixels. Finally, we proposed a white-box BREPS attack, which maintains the realism of optimized bboxes and efficiently finds adversarial prompts for the minimization and maximization of segmentation quality. We also formulated a robustness score and carried out a large-scale comparison of 15 models on 10 datasets from general to medical segmentation domains.

Acknowledgements

This work was supported by the The Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of Breast Ultrasound Images. *Data in Brief*, 28: 104863.

Antonov, A.; Moskalenko, A.; Shepelev, D.; Krapukhin, A.; Soshin, K.; Konushin, A.; and Shakhuro, V. 2024. RClicks: Realistic Click Simulation for Benchmarking Interactive Segmentation. *Advances in Neural Information Processing Systems*, 37: 127673–127710.

Arbeláez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2011. Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5): 898–916.

Bauchwitz, B.; and Cummings, M. 2025. Task configuration impacts annotation quality and model training performance in crowdsourced image segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6646–6656. IEEE.

Bernard, O.; Lalande, A.; et al. 2018. Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525.

Carlier, A.; Charvillat, V.; Salvador, A.; Giro-i Nieto, X.; and Marques, O. 2014. Click'n'Cut: Crowdsourced Interactive Segmentation with Object Candidates. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, CrowdMM '14, 53–56. New York, NY, USA: Association for Computing Machinery. ISBN 9781450331289.

Chen, W.-T.; Vong, Y.-J.; Kuo, S.-Y.; Ma, S.; and Wang, J. 2024. RobustSAM: segment anything robustly on degraded images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4081–4091.

Chen, X.; Zhao, Z.; Zhang, Y.; Duan, M.; Qi, D.; and Zhao, H. 2022. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1300–1309.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sam-med2d. *arXiv preprint arXiv:2308.16184*.

Croce, F.; and Hein, M. 2024. Segment (Almost) Nothing: Prompt-Agnostic Adversarial Attacks on Segmentation Models . In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 425–442. Los Alamitos, CA, USA: IEEE Computer Society.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 9, 48–59.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guan, Z.; Hu, M.; Zhou, Z.; Zhang, J.; Li, S.; and Liu, N. 2024. BadSAM: Exploring Security Vulnerabilities of SAM via Backdoor Attacks (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 23506–23507.

Huang, Y.; Yang, X.; Zhou, H.; Cao, Y.; Dou, H.; Dong, F.; and Ni, D. 2024. Robust box prompt based SAM for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 1–11. Springer.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2023. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, B.; Xiao, H.; and Tang, L. 2024. Asam: Boosting segment anything model with adversarial tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3699–3710.

Lin, T.-Y.; Maire, M.; Belongie, S.; et al. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755.

Liu, Q.; Xu, Z.; Bertasius, G.; and Niethammer, M. 2023a. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22290–22300.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Mazurowski, M. A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; and Zhang, Y. 2023. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89: 102918.

Moskalenko, A.; Shakhuro, V.; Vorontsova, A.; Konushin, A.; Antonov, A.; Krapukhin, A.; Shepelev, D.; and Soshin, K. 2024. TETRIS: towards exploring the robustness of interactive segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4287–4295.

Myers-Dean, J.; Fan, Y.; Price, B.; Chan, W.; and Gurari, D. 2024. Interactive Segmentation for Diverse Gesture Types Without Context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7198–7208.

Papadopoulos, D. P.; Uijlings, J. R.; Keller, F.; and Ferrari, V. 2017. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, 4930–4939.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *CVPR*.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*.

Rahman, M. M.; Munir, M.; Jha, D.; Bagci, U.; and Marculescu, R. 2024. PP-SAM: Perturbed Prompts for Robust Adaption of Segment Anything Model for Polyp Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4989–4995.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. In *ACM SIGGRAPH*, 309–314.

Sekachev, B.; Manovich, N.; Zhiltsov, M.; Zhavoronkov, A.; Kalinin, D.; Hoff, B.; TOsmanov; Kruchinin, D.; Zankevich, A.; DmitriySidnev; Markelov, M.; Johannes222; Chenuet, M.; a andre; telenachos; Melnikov, A.; Kim, J.; Ilouz, L.; Glazov, N.; Priya4607; Tehrani, R.; Jeong, S.; Skubriev, V.; Yonekura, S.; vugia truong; zliang7; lizhming; and Truong, T. 2020. opencv/cvat: v1.1.0.

Sofiiuk, K.; Petrov, I. A.; and Konushin, A. 2022. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3141–3145. IEEE.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing Annotations for Visual Object Detection. *HCOMP@ AAAI*, 1.

Supervisely. 2024. Supervisely: Computer Vision Annotation Platform. https://supervisely.com. Accessed: 2025-07-01.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020. Label Studio: Data Labeling Software. https://github.com/HumanSignal/label-studio. Accessed: 2025-07-01.

Toloka AI. 2025. Toloka Crowdsourcing Platform. https://toloka.ai. Accessed: 2025-07-01.

Tzutalin. 2015. LabelImg. https://github.com/tzutalin/labelImg. Accessed: 2025-07-01.

V7 Labs. 2024. V7 Darwin: Data Labeling Platform. https://www.v7labs.com/darwin. Accessed: 2025-07-01.

Wang, Y.; Zhao, Y.; and Petzold, L. 2024. An empirical study on the robustness of the segment anything model (sam). *Pattern Recognition*, 155: 110685.

Wong, H. E.; Rakic, M.; Guttag, J.; and Dalca, A. V. 2024. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image. *European Conference on Computer Vision (ECCV)*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023a. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Zhang, C.; Han, D.; Zheng, S.; Choi, J.; Kim, T.-H.; and Hong, C. S. 2023b. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579*.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing Through ADE20K Dataset. In *CVPR*.

Zhou, M.; Wang, H.; Zhao, Q.; Li, Y.; Huang, Y.; Meng, D.; and Zheng, Y. 2023. Interactive segmentation as gaussion process classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19488–19497.

Zhou, Z.; Song, Y.; Li, M.; Hu, S.; Wang, X.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024. Darksam: Fooling segment anything model to segment nothing. *Advances in Neural Information Processing Systems*, 37: 49859–49880.