How Resilient are Foundation Segmentation Models to Noisy Ultrasound Artifacts? A Multi-Dataset Benchmark Across Prompts, Noise Types, and Severity Levels

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Promptable SAM-family models are promising for interactive medical segmentation, but their robustness under ultrasound degradation remains under-characterized. We benchmark nine models on three public ultrasound datasets (BUSI, CAMUS, and DDTI) using 12 clinically motivated corruptions, clean L0 plus corrupted levels L1–L9, and three prompt modes: bounding box, point, and point+bbox. Corruption reduces mean Dice from 0.710 to 0.679 and increases mean HD95 from 51.41 to 54.91 pixels. MedSAM2 has the strongest absolute accuracy, yet in the dataset-balanced bbox table it drops from first on clean images to fourth under corruption; image-pooled rank stability assigns nonzero rank-1 probability only to SAM3 (0.77) and UltraSAM (0.23). Bbox prompting is the most reliable interaction mode, but perturbation analysis shows that its advantage depends strongly on box tightness. Pixelation and salt-and-pepper noise are most damaging. Overall, clean accuracy and corruption stability rank models differently, highlighting the need to evaluate model choice and prompt design under clinically plausible degradation.

Keywords: Segment Anything Model · Ultrasound · Robustness · Foundation models · Benchmark

1 Introduction

SAM [11] enabled promptable segmentation from sparse user input and has since expanded to video, medical, and concept-aware variants [5,13,14,20,26]. In medical imaging, these models are attractive because they can support interactive annotation and weakly supervised workflows without retraining a task-specific segmenter for every anatomy. Ultrasound, however, is highly sensitive to acquisition hardware, patient motion, gain, operator settings, and compression [4,9,24]. A model that performs well on clean benchmark images may therefore fail when image quality changes [6, 10, 22].

Prior SAM robustness work mainly uses natural-image corruptions or focuses on the original SAM [19, 21, 23], while medical SAM benchmarks emphasize

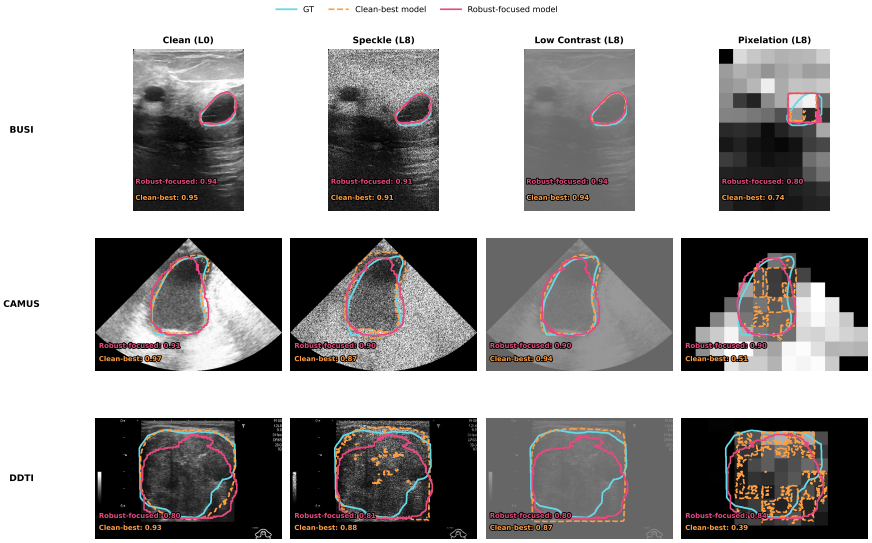

Fig. 1. Clean versus severely corrupted ultrasound under bbox prompts.

average accuracy rather than corruption stability [2, 8, 15, 16]. We provide a unified ultrasound benchmark across nine SAM-family models, three datasets, 12 corruption types, three prompt modes, and prompt perturbations. The key question is not only which model is most accurate, but which remains stable when ultrasound quality or prompts degrade. Our results show that these two criteria can lead to different model rankings.

2 Related Work

Medical SAM variants such as MedSAM [14], MedSAM2 [26], MedSAM3 [13], SAM-Med2D [7], UltraSAM [17], and MedicoSAM [3] adapt promptable segmentation to medical images, but they do not resolve robustness under degraded ultrasound. Existing medical evaluations also show that prompt type matters: bbox prompts often outperform point prompts, especially for irregular structures [8, 16]. Separately, robustness studies show sensitivity to blur, impulse noise, and compression [19, 21, 23, 25]. Our benchmark combines these axes, evaluating model family, dataset, prompt mode, corruption type, and severity within one controlled ultrasound protocol.

3 Methodology

Benchmark. Figure 2 summarizes the protocol. We evaluate 9 models on BUSI, CAMUS, and DDTI under bbox, point, and point+bbox prompts; 12 corruption types; clean L0 plus corrupted L1–L9. Unless stated otherwise, noisy results

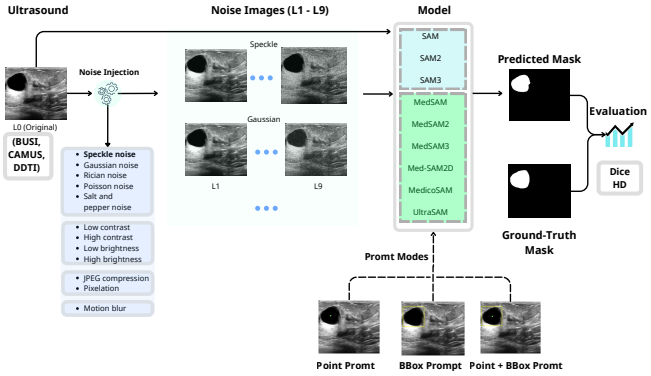

Fig. 2. Benchmark pipeline across models, prompts, corruptions, and severity levels.

average over L1–L9, while clean L0 is reported separately. This design lets us ask three different questions: which model is strongest on clean data, which prompt mode is most reliable, and which model degrades least when the input is corrupted. Main metrics are Dice and HD95.

Corruptions and prompts. We apply Gaussian, Rician, Poisson, salt-and-pepper, speckle, motion blur, low/high contrast, low/high brightness, JPEG compression, and pixelation to grayscale images in the original $[0, 255]$ space before model-specific preprocessing. Intensities are clipped after corruption; for speckle, $\tilde{I} = \text{clip}(I + I \cdot n)$ with $n \sim \mathcal{N}(0, \sigma^2)$. Severity schedules span Gaussian/Rician σ 5–80/5–75, Poisson peak 80–4, salt-and-pepper amount 0.01–0.50, speckle σ 0.05–0.90, blur kernel 3–40, JPEG quality down to 5, and pixelation block size up to 50. The corruption families are clinically motivated, but the levels are controlled stress tests rather than estimates of clinical prevalence: L1–L6 represent mild-to-moderate degradation, while L7–L9 deliberately probe out-of-distribution failure modes.

Prompts are derived from ground truth for controlled evaluation: bbox uses a tight mask box expanded by 5% per side, point uses one interior foreground point, and point+bbox provides both. This isolates corruption effects but is optimistic; Sect. 4.3 tests prompt uncertainty. Dice is $2|\hat{M} \cap M|/(|\hat{M}| + |M|)$. HD95 is the 95th percentile symmetric boundary distance, reported in pixels. BUSI and DDTI do not provide calibrated spacing, so cross-dataset HD95 comparisons are qualitative; CAMUS spacing enables millimetre conversion in the code/data package. For models returning multiple masks, the selected mask is the one with highest model-predicted IoU/confidence; ground truth is used only for evaluation.

Data, models, and statistics. BUSI contributes 647 abnormal breast ultrasound images; CAMUS contributes 2,000 echocardiography frames from 500 patients, using only the LV endocardial cavity; DDTI contributes 637 thyroid images [1,

12,18]. Corruptions affect inputs only, not masks, so performance changes reflect image degradation rather than label changes. All nine models use published ViT-B-scale checkpoints and official preprocessing; no model is fine-tuned for this benchmark. Reproducibility uses base seed 42 plus a hash of dataset, image id, corruption, and severity. Uncertainty uses paired cluster bootstrap with 2,000 resamples: image-level for BUSI/DDTI and patient-level for CAMUS, whose frames are correlated. Leaderboards are dataset-balanced so CAMUS does not dominate the pooled mean; rank stability is image-pooled, matching the Q6 bootstrap analysis.

4 Results

Overall, corruption reduces mean Dice from 0.7100 to 0.6792 (−4.33%) and increases mean HD95 from 51.41 to 54.91 pixels (+6.81%). The average change is modest, but model-level and prompt-level analyses reveal much sharper differences.

4.1 Absolute Quality and Robustness Differ

Under bbox prompting, MedSAM2 has the best clean Dice (0.9031) and clean HD95 (14.31 px) but the largest degradation (ΔDice 0.0458), falling from first on clean images to fourth under corruption (Table 1). Image-pooled rank stability assigns nonzero rank-1 probability only to SAM3 (0.766) and UltraSAM (0.234); MedSAM2's modal noisy rank is third. UltraSAM is nearly flat in dataset-balanced Dice, although its interpretation is complicated by training-data exposure (Sect. 5). Thus, clean accuracy and robustness are related but not interchangeable: MedSAM2 is best when the image is clean and the prompt is strong, whereas SAM3 and UltraSAM are more stable under corrupted bbox prompting.

Figure 3 visualizes this separation at the model level. The models with the highest clean Dice remain strong choices when images are clean, but the slope between clean and corrupted performance reveals a different deployment property: stability under image-quality shift. This is why we report both absolute Dice/HD95 and degradation-based quantities rather than collapsing the benchmark into a single leaderboard.

4.2 Prompting Drives Robustness

Bbox prompts are strongest and most stable: Dice is 0.855 clean and 0.838 noisy, compared with 0.754/0.739 for point+bbox and 0.520/0.460 for point-only. HD95 follows the same ordering, rising from 25.44 to 27.66 px for bbox and from 88.98 to 96.20 px for point-only. This shows that prompt modality is not only an interface detail; it determines how much spatial prior the model can use when image cues degrade. Some prompt modes are model-dependent: MedSAM is box-oriented, and UltraSAM performs similarly for point-only and

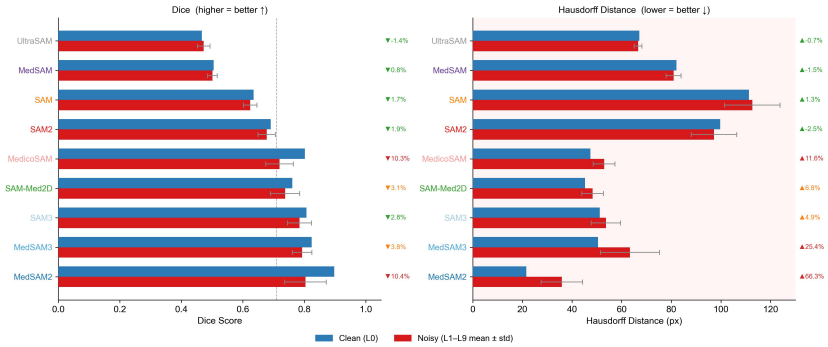

Fig. 3. Model-wise clean and corrupted performance. Absolute accuracy and degradation under corruption do not induce the same ranking.

Table 1. Dataset-balanced bbox leaderboard, ordered by noisy Dice. ΔDice is clean minus noisy; Severe Retention is Severe (L7–L9) Dice over clean Dice. †Trained on all evaluation datasets.

Model	Clean Dice	Noisy Dice	ΔDice	Severe Ret.	Clean HD95	Noisy HD95
SAM3	0.8804	**0.8710**	+0.0094	0.980	21.17	**21.81**
UltraSAM†	0.8638	0.8641	**−0.0003**	**1.000**	24.25	24.17
MedSAM3	0.8753	0.8607	+0.0145	0.968	22.49	24.75
MedSAM2	**0.9031**	0.8573	+0.0458	0.918	**14.31**	22.07
SAM-Med2D	0.8534	0.8337	+0.0197	0.958	26.26	28.31
MedicoSAM	0.8727	0.8321	+0.0406	0.939	22.81	27.47
MedSAM	0.8271	0.8250	+0.0021	0.994	27.25	27.04
SAM2	0.8222	0.8060	+0.0162	0.963	35.40	37.53
SAM	0.7990	0.7943	+0.0047	0.979	35.05	35.76

point+bbox in our wrapper/configuration, so bbox is the fairest cross-model comparison.

4.3 Prompt Perturbation Sensitivity

Because default prompts use ground truth, we perturb them directly (Table 2). Box size error dominates: a 20% expansion drops Dice from 0.8394 to 0.6772, much more than a 10% shift or shrink. Point centering barely matters, but a random interior point drops Dice by 0.0356, suggesting that point failures arise when the point approaches ambiguous boundaries rather than from the specific centering rule. Thus, bbox prompting is powerful but box-quality dependent, and our default prompt setting should be read as an upper-bound interaction scenario rather than a fully realistic clinical workflow.

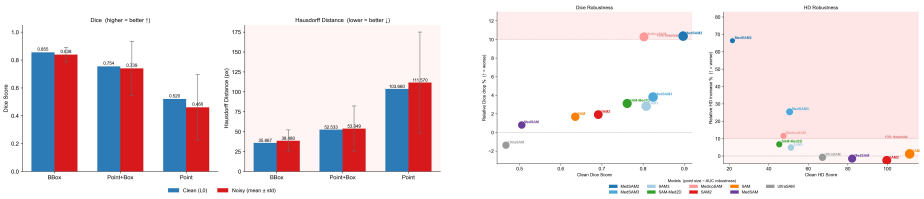

Fig. 4. Prompt mode effects and clean performance versus degradation.

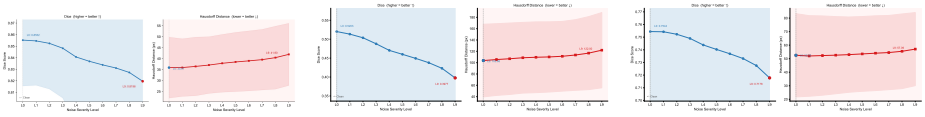

Fig. 5. Dice across L0–L9 under bbox, point, and point+bbox prompting.

4.4 Corruption and Dataset Effects

Corruption types differ sharply (Fig. 6). Pixelation is most damaging, followed by salt-and-pepper noise; JPEG compression, low contrast, and low brightness are milder at low severity. Motion blur harms boundaries more than Dice alone suggests, while synthetic speckle is not the worst case. A plausible explanation is that these models rely heavily on local boundary continuity; corruptions that destroy structure or introduce impulse artifacts are therefore more harmful than texture-like multiplicative noise.

Pooled averages hide anatomy-specific ranking flips (Table 3). UltraSAM is seventh on BUSI but first on CAMUS and DDTI; MedicoSAM is second on BUSI but last on CAMUS; SAM3 alone stays in the top three on every dataset. Boundary errors show similar anatomy dependence, with SAM2's CAMUS HD95 much higher than its BUSI/DDTI values. Since CAMUS contributes 2,000 of 3,284 images, per-dataset results are more informative than a single pooled leaderboard.

Figure 7 gives the complementary all-prompts view behind the bbox table. BUSI generally yields higher Dice for several medical SAM variants, whereas CAMUS and DDTI show tighter margins and more visible rank reversals. These differences support treating dataset identity as part of the robustness question rather than only as a source of more images.

5 Discussion

The benchmark shows that clean accuracy is an incomplete proxy for robustness. MedSAM2 maximizes clean Dice, but SAM3 gives the best accuracy–stability balance and is most uniform across anatomies. This distinction matters for deployment: a model selected only from clean-image Dice may not be the best choice once ultrasound quality varies. The same lesson applies to interaction design. Bbox workflows should be preferred when available, but boxes should be

Table 2. Prompt perturbation sensitivity over L0/L3/L5/L7 × 12 corruptions × 9 models × 3 datasets. Intervals are 95% cluster bootstrap.

Prompt variant	Dice [95% CI]	ΔDice [95% CI]
bbox default (expand 5%)	**0.8394 [0.8375, 0.8412]**	—
bbox expand 10%	0.7964 [0.7940, 0.7986]	−0.0430 [−0.0442, −0.0418]
bbox expand 20%	0.6772 [0.6742, 0.6803]	−0.1622 [−0.1645, −0.1598]
bbox shift 10%	0.8146 [0.8127, 0.8165]	−0.0248 [−0.0253, −0.0242]
bbox shrink 10%	0.8260 [0.8246, 0.8273]	−0.0134 [−0.0146, −0.0122]
point default	**0.4606 [0.4568, 0.4646]**	—
point centroid	0.4604 [0.4566, 0.4644]	−0.0001 [−0.0003, −0.0000]
point box-center	0.4603 [0.4562, 0.4642]	−0.0003 [−0.0009, 0.0002]
point random inside	0.4251 [0.4211, 0.4291]	−0.0356 [−0.0367, −0.0344]

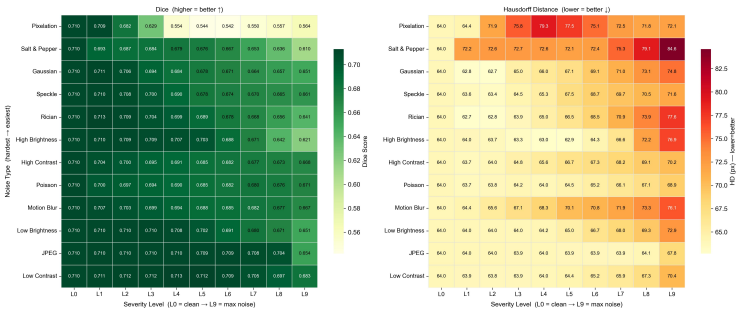

Fig. 6. Dice and HD95 by corruption type and severity.

tight: a 20% over-draw costs more Dice than a 10% translation. Point-only workflows are higher risk because they degrade under both corruption and prompt imprecision.

Training exposure complicates interpretation. SAM, SAM2, and SAM3 are trained on natural images and act as zero-shot controls; UltraSAM explicitly includes BUSI, CAMUS, and DDTI in training, so its flat degradation and CAMUS/DDTI leads carry an in-distribution caveat. Yet UltraSAM ranks only seventh on BUSI, so exposure alone does not explain all behavior. For other medical SAM variants, incomplete training manifests mean image-level overlap cannot be excluded. We therefore interpret the benchmark as a practical robustness comparison, not a causal attribution of robustness to architecture or training data.

This is a promptable SAM-family robustness benchmark, not a comparison with task-specific supervised systems such as nnU-Net or U-Net variants, and absolute Dice should not be read as ultrasound SOTA. The results are most useful for choosing among promptable foundation models and for designing robustness evaluations. Limitations include GT-derived prompts, synthetic rather

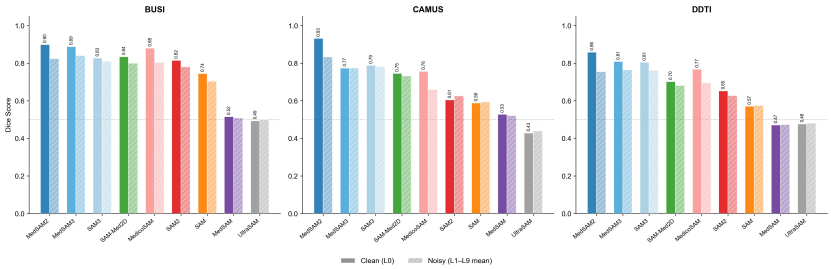

Fig. 7. Per-dataset clean and corrupted Dice across BUSI, CAMUS, and DDTI. Dataset-specific margins expose ranking changes hidden by pooled means.

Table 3. Per-dataset noisy bbox performance; best per column in bold. †Trained on all evaluation datasets.

	Noisy Dice			Noisy HD95 (px)		
Model	**BUSI**	**CAMUS**	**DDTI**	**BUSI**	**CAMUS**	**DDTI**
SAM3	0.8708	0.8734	0.8689	**23.26**	24.14	**18.02**
UltraSAM†	0.8406	**0.8813**	**0.8705**	30.14	23.34	19.01
MedSAM3	**0.8788**	0.8445	0.8591	24.44	29.82	19.98
MedSAM2	0.8637	0.8542	0.8542	25.48	**20.89**	19.83
SAM-Med2D	0.8531	0.8267	0.8215	27.81	32.02	25.07
MedicoSAM	0.8750	0.7776	0.8439	25.66	34.66	22.05
MedSAM	0.8208	0.8555	0.7989	29.67	26.63	24.79
SAM2	0.8410	0.7848	0.7925	29.86	49.62	33.12
SAM	0.8193	0.7891	0.7747	36.14	40.34	30.75

than clinically calibrated corruptions, 2D static ultrasound only, no downstream diagnostic or clinician workflow evaluation, and no confidence-calibrated failure detector because model-predicted IoU scores were not persisted. Future work should validate the same protocol on real degraded acquisitions, clinician-drawn prompts, video/3D ultrasound, and downstream measurement error.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020)
2. Ali, M., et al.: A review of the segment anything model (sam) for medical image analysis: Accomplishments and perspectives. Computerized Medical Imaging and Graphics **119**, 102473 (2025)
3. Archit, A., Freckmann, L., Pape, C.: Medicosam: Robust improvement of sam for medical imaging. IEEE Transactions on Medical Imaging (2025)
4. Baad, M., Lu, Z.F., Reiser, I., Paushter, D.: Clinical significance of us artifacts. RadioGraphics **37**(5), 1408–1423 (2017). https://doi.org/10.1148/rg.2017160175, `https://doi.org/10.1148/rg.2017160175`, pMID: 28777700

5. Carion, N., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chen, E., Mathai, T., Sarode, V., Choset, H., Galeotti, J.M.: A study of domain generalization on ultrasound-based multi-class segmentation of arteries, veins, ligaments, and nerves using transfer learning. ArXiv **abs/2011.07019** (2020), https://api.semanticscholar.org/CorpusID:226956106
7. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023), https://arxiv.org/abs/2308.16184
8. Dong, H., Gu, H., Chen, Y., Yang, J., Chen, Y., Mazurowski, M.A.: Segment anything model 2: An application to 2d and 3d medical images. IEEE Transactions on Biomedical Engineering (2026)
9. Feldman, M.K., Katyal, S., Blackwood, M.S.: Us artifacts. RadioGraphics **29**(4), 1179–1189 (2009). https://doi.org/10.1148/rg.294085199, https://doi.org/10.1148/rg.294085199, pMID: 19605664
10. Jiang, J., Jiang, X., Xu, L., Zhang, Y., Zheng, Y., Kong, D.: Noise-robustness test for ultrasound breast nodule neural network models as medical devices. Frontiers in Oncology **Volume 13 - 2023** (2023). https://doi.org/10.3389/fonc.2023.1177225, https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2023.1177225
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
12. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
13. Liu, A., et al.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Ma, Jun and others: Segment anything in medical images and videos: Benchmark and deployment (2024). https://doi.org/10.48550/arXiv.2408.03322
16. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023)
17. Meyer, A., Murali, A., Zarin, F., Mutter, D., Padoy, N.: Ultrasam: A foundation model for ultrasound using large open-access segmentation datasets. International Journal of Computer Assisted Radiology and Surgery **21**(1), 93–102 (Sep 2025). https://doi.org/10.1007/s11548-025-03517-8, http://dx.doi.org/10.1007/s11548-025-03517-8
18. Pedraza, L., Vargas, C., Narvaez, F., Duran, O., Munoz, E., Romero, E.: An open access thyroid ultrasound-image database. In: Proceedings of SPIE (2015)
19. Qiao, Y., Zhang, C., Kang, T., Kim, D., Zhang, C., Hong, C.S.: Robustness of sam: Segment anything under corruptions and beyond. arXiv preprint arXiv:2306.07713 (2023)
20. Ravi, N., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
21. Schiappa, M.C., et al.: Robustness analysis on foundational segmentation models. arXiv preprint arXiv:2306.09278 (2023)

22. Sun, J., Li, C., Lu, Z., He, M., Zhao, T., Li, X., Gao, L., Xie, K., Lin, T., Sui, J., Xi, Q., Zhang, F., Ni, X.: Tnsnet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision. Computer Methods and Programs in Biomedicine **215**, 106600 (2022). https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106600, https://www.sciencedirect.com/science/article/pii/S016926072100674X
23. Wang, Y., Zhao, Y., Petzold, L.: An empirical study on the robustness of the segment anything model (sam). Pattern Recognition **155**, 110685 (2024)
24. Yang, Y., Chen, Y., Dong, X., Zhang, J., Long, C., Jin, Z., Dai, Y.: An annotated heterogeneous ultrasound database. Scientific Data **12**(1), 148 (Jan 2025). https://doi.org/10.1038/s41597-025-04464-4, https://doi.org/10.1038/s41597-025-04464-4
25. Yu, J., et al.: Sam 2 in robotic surgery: An empirical evaluation for robustness and generalization in surgical video segmentation. arXiv preprint arXiv:2408.04593 (2024)
26. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)