This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material When Does Prompt-Perturbation Uncertainty Catch Interactive-Segmentation Failures?

Sathvik Loke[1], Neeraj Sai Venkat Movva[1], Mishka Hota[2], Rishith Gudepu[1], Nelson Mburu[1], and Suhrit Bestavemula[2]

[1] Illinois Mathematics and Science Academy, Aurora, IL, USA
sloke@imsa.edu
[2] Neuqua Valley High School, Naperville, IL, USA

1 Full paired-comparison detail (Table S1)

The main paper reports Holm-corrected significance as stars on the headline tables. Table S1 gives the underlying numbers for every comparison: the effect size, the uncorrected p-value from each test, and the corrected value. We report both the parametric and the resampling test because they disagree in exactly the situation the paper is most cautious about, namely cells with very few catastrophic failures, where the normal approximation behind DeLong's test is doing the most work.

2 Threshold sweep (Table S2)

The catastrophic threshold (Dice < 0.3) is the operating point a triage system cares about, and is what the main paper reports. Table S2 gives the full sweep.

3 Error-target convention (Table S3)

Section 3.4 of the main paper notes that the one-shot baselines derive from a different mask than the consensus, so there is a choice about which mask's error they should be asked to predict. Table S3 reports the localization results under both conventions.

4 Realistic prompts, full tables (Tables S4–S6)

Tables S4 and S5 repeat the two headline tables under the realistic seed-prompt regime, and Table S6 gives the change relative to oracle prompts, from which the summary in the main paper is drawn.

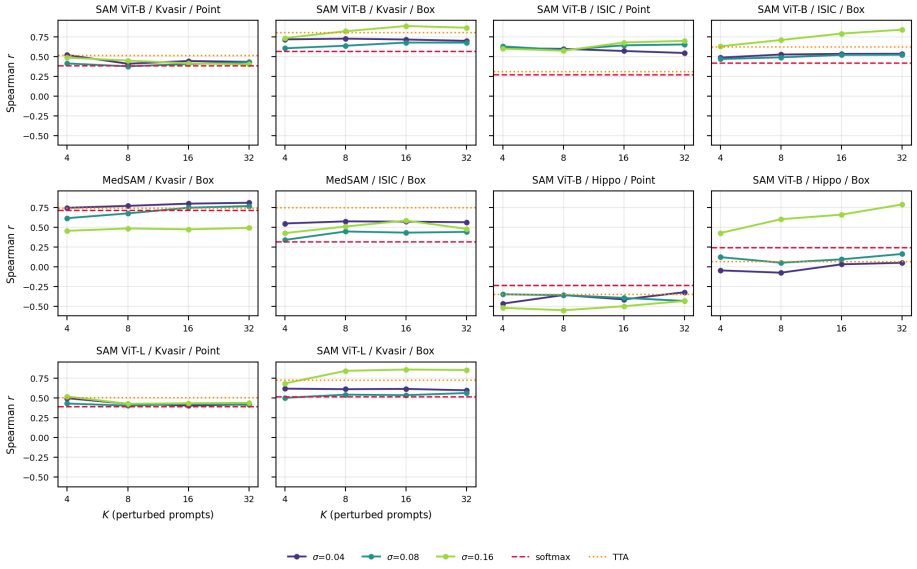

Fig. S1. PPU's rank correlation with error as a function of the number of perturbed prompts K, at three jitter scales, for each cell. Dashed line: softmax entropy. Dotted line: TTA. Both baselines are evaluated at the default setting and do not depend on K.

5 Cost breakdown (Table S7)

6 Localization (Table S8)

7 Hyper-parameter sweep (Fig. S1)

8 Qualitative example

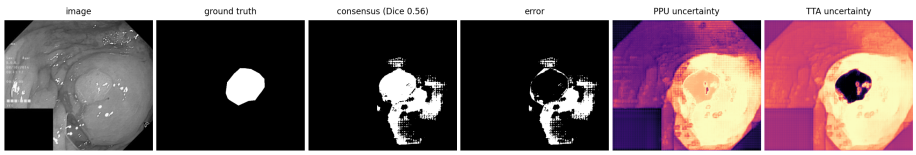

Fig. S2. A partial failure on Kvasir-SEG under a point prompt. Left to right: image, ground truth, the PPU consensus mask, its error against ground truth, and the PPU and TTA uncertainty maps. The consensus has leaked well beyond the polyp. Both uncertainty maps are scaled to their own 1st–99th percentile, since their absolute magnitudes are not comparable and only the spatial pattern is. TTA is confidently certain over the region it segments correctly, while PPU spreads uncertainty across the whole leaked area.

Table S1. Every paired comparison of PPU against a baseline. Δ is PPU minus the baseline, so positive favours PPU. p_{DeLong} is the parametric test for correlated ROC curves and is defined for AUROC only; p_{perm} is the paired permutation test over 10^4 resamples. The last column is the Holm-corrected value within the metric's family, and is the one any claim should rest on.

Backbone	Dataset	Mode	Baseline	Metric	Δ	p_{DeLong}	p_{perm}	p_{Holm}
SAM ViT-B	Kvasir	Point	Softmax	AUROC	+0.096	0.500	0.507	1.000
SAM ViT-B	Kvasir	Point	Softmax	Spearman	+0.020	—	0.845	1.000
SAM ViT-B	Kvasir	Point	`iou_pred`	AUROC	+0.090	0.081	0.318	1.000
SAM ViT-B	Kvasir	Point	`iou_pred`	Spearman	+0.125	—	0.148	1.000
SAM ViT-B	Kvasir	Point	TTA	AUROC	−0.017	0.833	0.851	1.000
SAM ViT-B	Kvasir	Point	TTA	Spearman	−0.110	—	0.163	1.000
SAM ViT-B	Kvasir	Box	Softmax	AUROC	+0.014	0.450	0.492	1.000
SAM ViT-B	Kvasir	Box	Softmax	Spearman	+0.111	—	0.037	0.927
SAM ViT-B	Kvasir	Box	`iou_pred`	AUROC	+0.001	0.915	0.979	1.000
SAM ViT-B	Kvasir	Box	`iou_pred`	Spearman	−0.072	—	0.400	1.000
SAM ViT-B	Kvasir	Box	TTA	AUROC	+0.006	0.690	0.811	1.000
SAM ViT-B	Kvasir	Box	TTA	Spearman	−0.127	—	0.035	0.913
SAM ViT-B	ISIC	Point	Softmax	AUROC	+0.361	0.010	0.004	0.225
SAM ViT-B	ISIC	Point	Softmax	Spearman	+0.369	—	0.002	0.053
SAM ViT-B	ISIC	Point	`iou_pred`	AUROC	+0.256	0.016	0.089	0.362
SAM ViT-B	ISIC	Point	`iou_pred`	Spearman	+0.212	—	0.043	0.996
SAM ViT-B	ISIC	Point	TTA	AUROC	+0.193	0.080	0.096	1.000
SAM ViT-B	ISIC	Point	TTA	Spearman	+0.329	—	0.001	0.041
SAM ViT-B	ISIC	Box	Softmax	AUROC	+0.273	0.123	0.142	1.000
SAM ViT-B	ISIC	Box	Softmax	Spearman	+0.101	—	0.296	1.000
SAM ViT-B	ISIC	Box	`iou_pred`	AUROC	+0.078	0.107	0.684	1.000
SAM ViT-B	ISIC	Box	`iou_pred`	Spearman	−0.121	—	0.413	1.000
SAM ViT-B	ISIC	Box	TTA	AUROC	+0.060	0.062	0.621	1.000
SAM ViT-B	ISIC	Box	TTA	Spearman	−0.106	—	0.325	1.000
MedSAM	Kvasir	Box	Softmax	AUROC	−0.072	0.119	0.413	1.000
MedSAM	Kvasir	Box	Softmax	Spearman	+0.032	—	0.775	1.000
MedSAM	Kvasir	Box	`iou_pred`	AUROC	+0.251	<0.001	<0.001	0.003
MedSAM	Kvasir	Box	`iou_pred`	Spearman	+0.723	—	<0.001	0.003
MedSAM	Kvasir	Box	TTA	AUROC	−0.019	0.623	0.807	1.000
MedSAM	Kvasir	Box	TTA	Spearman	+0.005	—	0.953	1.000
MedSAM	ISIC	Box	Softmax	Spearman	+0.116	—	0.498	1.000
MedSAM	ISIC	Box	`iou_pred`	Spearman	+0.178	—	0.273	1.000
MedSAM	ISIC	Box	TTA	Spearman	−0.317	—	0.039	0.934
SAM ViT-B	Hippo	Point	Softmax	AUROC	−0.077	0.274	0.470	1.000
SAM ViT-B	Hippo	Point	Softmax	Spearman	−0.162	—	0.384	1.000
SAM ViT-B	Hippo	Point	`iou_pred`	AUROC	+0.037	0.502	0.551	1.000
SAM ViT-B	Hippo	Point	`iou_pred`	Spearman	−0.004	—	0.964	1.000
SAM ViT-B	Hippo	Point	TTA	AUROC	−0.017	0.789	0.789	1.000
SAM ViT-B	Hippo	Point	TTA	Spearman	−0.045	—	0.649	1.000
SAM ViT-B	Hippo	Box	Softmax	Spearman	−0.148	—	0.269	1.000
SAM ViT-B	Hippo	Box	`iou_pred`	Spearman	−0.162	—	0.374	1.000
SAM ViT-B	Hippo	Box	TTA	Spearman	+0.026	—	0.872	1.000
SAM ViT-L	Kvasir	Point	Softmax	AUROC	+0.217	0.289	0.294	1.000
SAM ViT-L	Kvasir	Point	Softmax	Spearman	+0.032	—	0.704	1.000
SAM ViT-L	Kvasir	Point	`iou_pred`	AUROC	+0.084	0.233	0.491	1.000
SAM ViT-L	Kvasir	Point	`iou_pred`	Spearman	+0.010	—	0.886	1.000
SAM ViT-L	Kvasir	Point	TTA	AUROC	+0.124	0.306	0.384	1.000
SAM ViT-L	Kvasir	Point	TTA	Spearman	−0.083	—	0.288	1.000
SAM ViT-L	Kvasir	Box	Softmax	AUROC	+0.024	0.385	0.628	1.000
SAM ViT-L	Kvasir	Box	Softmax	Spearman	+0.017	—	0.813	1.000
SAM ViT-L	Kvasir	Box	`iou_pred`	AUROC	+0.075	0.068	0.388	1.000
SAM ViT-L	Kvasir	Box	`iou_pred`	Spearman	−0.127	—	0.192	1.000

Table S2. Failure-flag AUROC across the full Dice-threshold sweep. The main paper reports the catastrophic threshold (0.3); this table shows that the qualitative pattern is monotone in the threshold, and that the number of positives grows as the threshold loosens.

Backbone	Dataset	Mode	Dice <	n_{fail}	PPU	Softmax	`iou_pred`	TTA
SAM ViT-B	Kvasir	Point	0.3	19	0.67	0.57	0.58	0.69
SAM ViT-B	Kvasir	Point	0.5	32	0.59	0.66	0.50	0.66
SAM ViT-B	Kvasir	Point	0.7	55	0.66	0.69	0.59	0.71
SAM ViT-B	Kvasir	Point	0.8	83	0.69	0.71	0.59	0.72
SAM ViT-B	Kvasir	Box	0.3	24	0.96	0.94	0.96	0.95
SAM ViT-B	Kvasir	Box	0.5	31	0.98	0.93	0.97	0.96
SAM ViT-B	Kvasir	Box	0.7	36	0.97	0.92	0.97	0.97
SAM ViT-B	Kvasir	Box	0.8	42	0.98	0.94	0.98	0.98
SAM ViT-B	ISIC	Point	0.3	9	0.95	0.59	0.69	0.75
SAM ViT-B	ISIC	Point	0.5	24	0.86	0.64	0.66	0.58
SAM ViT-B	ISIC	Point	0.7	38	0.85	0.66	0.73	0.69
SAM ViT-B	ISIC	Point	0.8	45	0.83	0.65	0.72	0.67
SAM ViT-B	ISIC	Box	0.3	4	0.99	0.72	0.92	0.93
SAM ViT-B	ISIC	Box	0.5	9	0.96	0.78	0.86	0.92
SAM ViT-B	ISIC	Box	0.7	16	0.95	0.74	0.85	0.89
SAM ViT-B	ISIC	Box	0.8	31	0.82	0.72	0.85	0.83
MedSAM	Kvasir	Box	0.3	30	0.79	0.86	0.54	0.81
MedSAM	Kvasir	Box	0.5	36	0.81	0.90	0.52	0.81
MedSAM	Kvasir	Box	0.7	67	0.88	0.86	0.60	0.86
MedSAM	Kvasir	Box	0.8	88	0.90	0.87	0.53	0.88
MedSAM	ISIC	Box	0.3	0	—	—	—	—
MedSAM	ISIC	Box	0.5	0	—	—	—	—
MedSAM	ISIC	Box	0.7	3	0.89	0.83	0.52	0.98
MedSAM	ISIC	Box	0.8	9	0.76	0.68	0.62	0.97
SAM ViT-B	Hippo	Point	0.3	44	0.28	0.36	0.25	0.30
SAM ViT-B	Hippo	Point	0.5	100	—	—	—	—
SAM ViT-B	Hippo	Point	0.7	100	—	—	—	—
SAM ViT-B	Hippo	Point	0.8	100	—	—	—	—
SAM ViT-B	Hippo	Box	0.3	0	—	—	—	—
SAM ViT-B	Hippo	Box	0.5	1	1.00	0.62	0.94	0.84
SAM ViT-B	Hippo	Box	0.7	52	0.57	0.62	0.63	0.55
SAM ViT-B	Hippo	Box	0.8	92	0.38	0.69	0.46	0.37
SAM ViT-L	Kvasir	Point	0.3	8	0.67	0.45	0.59	0.55
SAM ViT-L	Kvasir	Point	0.5	16	0.64	0.65	0.64	0.61
SAM ViT-L	Kvasir	Point	0.7	33	0.70	0.73	0.66	0.70
SAM ViT-L	Kvasir	Point	0.8	55	0.70	0.72	0.67	0.69
SAM ViT-L	Kvasir	Box	0.3	20	0.93	0.91	0.86	0.92
SAM ViT-L	Kvasir	Box	0.5	23	0.94	0.90	0.87	0.92
SAM ViT-L	Kvasir	Box	0.7	30	0.96	0.91	0.86	0.91
SAM ViT-L	Kvasir	Box	0.8	33	0.95	0.91	0.87	0.91

Table S3. Localization AUROC under both error-target conventions. *common* scores every map against the error of the PPU consensus mask, so all methods are ranked on an identical error set; *own* scores each map against the error of the mask that method would itself return. PPU is unchanged by construction, since the common target *is* its own. The comparison matters because a method whose consensus happens to be more accurate is handed an easier ranking problem under the *own* convention.

Backbone	Dataset	Mode	PPU		Softmax		TTA	
			common	own	common	own	common	own
SAM ViT-B	Kvasir	Point	0.677	—	0.723	0.751	0.747	0.790
SAM ViT-B	Kvasir	Box	0.689	—	0.763	0.749	0.764	0.762
SAM ViT-B	ISIC	Point	0.666	—	0.706	0.702	0.720	0.731
SAM ViT-B	ISIC	Box	0.662	—	0.756	0.722	0.779	0.749
MedSAM	Kvasir	Box	0.645	—	0.719	0.686	0.727	0.708
MedSAM	ISIC	Box	0.823	—	0.890	0.878	0.901	0.898
SAM ViT-B	Hippo	Point	0.515	—	0.503	0.513	0.521	0.492
SAM ViT-B	Hippo	Box	0.841	—	0.818	0.856	0.784	0.848
SAM ViT-L	Kvasir	Point	0.673	—	0.757	0.809	0.768	0.823
SAM ViT-L	Kvasir	Box	0.671	—	0.778	0.776	0.796	0.791

Table S4. Catastrophic-failure flagging (AUROC at Dice < 0.3) with 95% bootstrap CIs. `iou_pred` is the model's own self-rated mask-quality score (1−score as uncertainty); TTA is dihedral test-time augmentation. **Bold** = best in row. Stars denote a paired DeLong test against PPU, Holm-corrected across all cells and baselines ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Their absence means the difference from PPU is *not* statistically distinguishable, which is the case for most rows. Cells with n_{fail}=0 leave the AUROC undefined. Seed prompts here are the realistic ones: off-centroid interior clicks and loosely drawn boxes.

Backbone	Dataset (N)	Mode	n_{fail}	PPU	Softmax	`iou_pred`	TTA
SAM ViT-B	Kvasir (200)	Point	24	**.77** [.64, .89]	.55 [.40, .69]	.64 [.49, .78]	.69 [.57, .81]
SAM ViT-B	Kvasir (200)	Box	30	**.95** [.91, .98]	.87 [.80, .93]	.91 [.86, .95]	.89 [.83, .94]
SAM ViT-B	ISIC (100)	Point	15	**.85** [.71, .97]	.54 [.35, .73]	.71 [.58, .84]	.65 [.47, .81]
SAM ViT-B	ISIC (100)	Box	10	**.90** [.82, .97]	.83 [.71, .93]	.80 [.54, .99]	.84 [.70, .94]
MedSAM	Kvasir (200)	Box	35	.75 [.66, .84]	**.79** [.71, .87]	.49 [.39, .59]**	.67 [.58, .75]
MedSAM	ISIC (100)	Box	0	—	—	—	—
SAM ViT-B	Hippo (100)	Point	40	.36 [.26, .47]	.36 [.26, .47]	.36 [.26, .47]	**.44** [.33, .56]
SAM ViT-B	Hippo (100)	Box	0	—	—	—	—
SAM ViT-L	Kvasir (200)	Point	14	**.93** [.85, .98]	.66 [.47, .83]	.71 [.54, .86]*	.76 [.60, .89]
SAM ViT-L	Kvasir (200)	Box	23	**.91** [.86, .96]	.84 [.74, .91]	.71 [.57, .83]	.82 [.71, .90]

Table S5. Spearman r of uncertainty vs. error (1−Dice) with 95% bootstrap CIs, computed over all cases rather than only the failures, so it uses the whole grid rather than the small positive class. **Bold** = best in row; stars denote a paired permutation test against PPU, Holm-corrected within this metric.

Backbone	Dataset	Mode	PPU	Softmax	`iou_pred`	TTA
SAM ViT-B	Kvasir	Point	+.39 [+.25, +.51]	+.38 [+.25, +.51]	+.39 [+.24, +.53]	+.**49** [+.37, +.61]
SAM ViT-B	Kvasir	Box	+.**75** [+.66, +.81]	+.64 [+.54, +.71]	+.66 [+.57, +.74]	+.72 [+.63, +.79]
SAM ViT-B	ISIC	Point	+.**55** [+.37, +.68]	+.27 [+.08, +.46]	+.46 [+.30, +.62]	+.33 [+.14, +.51]
SAM ViT-B	ISIC	Box	+.58 [+.44, +.70]	+.**59** [+.46, +.69]	+.39 [+.17, +.59]	+.45 [+.28, +.61]
MedSAM	Kvasir	Box	+.66 [+.56, +.73]	+.**67** [+.58, +.75]	−.10 [−.24, +.04]**	+.57 [+.46, +.66]
MedSAM	ISIC	Box	+.35 [+.17, +.51]	+.32 [+.12, +.50]	−.20 [−.39, −.01]*	+.**45** [+.28, +.59]
SAM ViT-B	Hippo	Point	−.26 [−.45, −.05]	−.24 [−.42, −.03]	−.23 [−.42, −.03]	−.**16** [−.35, +.06]
SAM ViT-B	Hippo	Box	−.34 [−.54, −.14]	+.**39** [+.21, +.55]**	−.11 [−.32, +.08]	−.15 [−.34, +.07]
SAM ViT-L	Kvasir	Point	+.34 [+.19, +.47]	+.37 [+.23, +.48]	+.44 [+.32, +.55]	+.**54** [+.43, +.65]
SAM ViT-L	Kvasir	Box	+.**62** [+.51, +.70]	+.51 [+.39, +.61]	+.44 [+.30, +.57]	+.60 [+.48, +.70]

Table S6. Effect of replacing oracle prompts with realistic ones. Δ is realistic minus oracle, so negative values mean the signal degrades once the prompt is no longer an oracle; values at or below −0.20 are set in bold. Mean Dice falls in every cell because the prompt is worse, which is expected. The question is whether the *uncertainty* stays informative about the resulting error.

Backbone	Dataset	Mode	Mean Dice		Δ Spearman (realistic − oracle)			
			oracle	real.	PPU	Softmax	`iou_pred`	TTA
SAM ViT-B	Kvasir	Point	0.744	0.713	−0.02	−0.00	+0.10	−0.02
SAM ViT-B	Kvasir	Box	0.791	0.743	+0.07	+0.07	−0.09	−0.08
SAM ViT-B	ISIC	Point	0.692	0.663	−0.09	−0.00	+0.03	+0.02
SAM ViT-B	ISIC	Box	0.802	0.721	+0.06	+0.16	**−0.25**	−0.17
MedSAM	Kvasir	Box	0.706	0.650	−0.09	−0.04	−0.13	−0.18
MedSAM	ISIC	Box	0.894	0.857	−0.08	+0.01	**−0.45**	**−0.30**
SAM ViT-B	Hippo	Point	0.311	0.311	+0.13	−0.01	+0.16	+0.19
SAM ViT-B	Hippo	Box	0.692	0.633	**−0.44**	+0.15	**−0.37**	**−0.21**
SAM ViT-L	Kvasir	Point	0.823	0.800	−0.09	−0.03	+0.03	+0.04
SAM ViT-L	Kvasir	Box	0.820	0.792	+0.08	−0.01	**−0.22**	−0.13

Table S7. Measured wall-clock cost per case for vit-b on one Apple M4 Pro (MPS), averaged over 15 cases with nothing else on the device. One encoder pass costs 1.19 s and one decoder pass 45 ms, a ratio of roughly 26:1. That ratio is the whole argument for perturbing the prompt instead of the image: PPU buys K samples for one encoder pass, while TTA pays a fresh encoder pass per augmentation and ends up 5.2× slower than PPU.

Method	Encoder passes	Decoder passes	Time (s)	Overhead
Single prediction (softmax, `iou_pred`)	1	1	1.63	1.0×
PPU (K=16)	1	16	2.12	1.3×
TTA (D_4, 8 transforms)	8	8	11.07	6.8×

Table S8. Pixel-level error localization (K=16). *Loc. AUROC* ranks every voxel in a dilated ROI around the object by uncertainty, so trivially correct background does not inflate the score. Precision@k measures how often the top-k uncertain voxels (with k matched to the number of true error voxels) are actually errors, reported directly and as a multiplier over the chance rate. **Bold** = best in row.

Backbone	Dataset (N)	Mode	Loc. AUROC			PPU prec@k	
			PPU	Softmax	TTA	value	(×chance)
SAM ViT-B	Kvasir (200)	Point	0.677	0.723	**0.747**	0.423	3.7×
SAM ViT-B	Kvasir (200)	Box	0.689	0.763	**0.764**	0.413	6.0×
SAM ViT-B	ISIC (100)	Point	0.666	0.706	**0.720**	0.493	2.7×
SAM ViT-B	ISIC (100)	Box	0.662	0.756	**0.779**	0.421	4.2×
MedSAM	Kvasir (200)	Box	0.645	0.719	**0.727**	0.432	3.9×
MedSAM	ISIC (100)	Box	0.823	0.890	**0.901**	0.416	7.0×
SAM ViT-B	Hippo (100)	Point	0.515	0.503	**0.521**	0.693	1.0×
SAM ViT-B	Hippo (100)	Box	**0.841**	0.818	0.784	0.453	4.5×
SAM ViT-L	Kvasir (200)	Point	0.673	0.757	**0.768**	0.399	5.4×
SAM ViT-L	Kvasir (200)	Box	0.671	0.778	**0.796**	0.403	6.1×